Enhancing Loan Approval Predictions: Comparative Analysis of Five Machine Learning Algorithms on Varied Datasets

Diaa Salama AbdElminaam[1], Asmaa Maher Khalifa[2],
Minna Hany Abdelhady[3], Mohamed Hisham Sayed [4], Nermien Mohamed Yaekoub[5],Toka Saeed Mohamed [6]
Faculty of Computer Science
Misr International University, Cairo, Egypt
diaa.salama[1], asmaa2110434[2],
minna2109122[3], mohamed2106505[4], nermien2110633[5], toka2104550 [6]{@miuegypt.edu.eg}

Enhancing Loan Approval Predictions: Comparative Analysis of Five Machine Learning Algorithms on Varied Data sets ***Abstract*—Their is a great significance of financial decision-making, loan approval prediction. As it has crucial effects on both the person and the whole economy.Given the critical role that loans play in supporting a variety of financial attempts, the importance of rapid loan approvals has increased in recent years.An individual's capacity to engage in traditional economic activities and maintain financial stability may be severely impacted by delays in the approval procedure.While it also includes problem statements such as Fairness and Bias Difficulties, in addition to the deficiency of forthcoming, In addition to dynamic finance conditions, the prediction model will encounter issues of this nature because of the swift alterations in the business community. Our objective in this paper to propose a new framework for loan prediction using machine learning. We applied this prediction on three data sets. Moreover, We compared between them using various algorithms and they are Decision Tree, Naive Bayes, KNN, Random Forest, Gradient Booting classifier.The result of the testing we concluded that gradient boosting classifier algorithm is the most suitable because its accuracy for the first data set (loan-approval-prediction) is 90% and for the second is 90% and for the third is 80%**

Keywords: Loan approval prediction; Machine Learning; Classification; Random Forest; K-Nearest Neighbor; Decision Tree; Support vector machine; Naive Bayes.

I. INTRODUCTION

The financial sector has witnessed a transformational shift in its operations in recent years, mainly to impressive technological developments and the widespread use of machine learning techniques. This paradigm change has left a permanent effect on the landscape of loan approval predictions, an important field in the financial sector. old ways for evaluating creditworthiness are going through significant modification, with the integration and, in certain situations, the replacement of these old approaches with extremely effective machine learning algorithms. These advanced algorithms have a remarkable ability to examine large data-sets, resulting in more precise and efficient granting choices. The combination of traditional credit assessment methodologies with cutting-edge machine learning technology is an important turning point in the financial industry's search for more accuracy, efficiency, and adaptability in loan approval processes.

Fig. 1. Loan prediction

In the current economic environment, bank and financial institution loan failures have become a major challenge. These failures happen when borrowers don't pay back their loans, which has a negative impact on lenders and the financial system as a whole. One of the main causes of loan failures is the insufficient evaluation of borrowers' financial standing, which results in credit being given to people or companies who have low repayment capacity. Furthermore, borrowers' financial stability may be severely impacted by unanticipated events and economic downturns, making it more challenging for them to fulfill their repayment commitments. Furthermore, insufficient surveillance and assessment of loan portfolios, among other risk management techniques within financial institutions, can make the issue worse. In addition to causing banks and other organizations to suffer financial losses, loan failures damage their reputation and reduce investor trust. In order to lessen the negative impacts of loan failures and protect the stability of the financial industry, these failures require the enactment of strong risk control frameworks, increased verification procedures, and enhanced oversight techniques.

Fig. 2. Showing the problem of loan prediction

This study seeks to give a comprehensive and detailed evaluation of five basic machine learning algorithms chosen for their relevance, performance measures, and widespread use across different data-sets. The algorithms chosen for evaluation have established themselves as significant participants in the field of loan approval predictions, demonstrating their effectiveness in dealing with complicated financial data. The comparative study conducted during this research will go into a complete investigation of the strengths and limits included in each algorithm, demonstrating their performance differences in various circumstances. The insights gained from this exhaustive examination will provide financial institutions with the knowledge they need to make reasonable and well-informed judgments about the adoption and use of these advanced technologies in loan approval procedures.The goal of this comparison analysis is to add to the ongoing discussion about enhancing loan approval forecasts by throwing light on the specific characteristics and applicability of each machine learning method.

Fig. 3. Loan Aprroval

This study examines categorization methods to predict the likelihood that a loan will be approved in financial institutions. The introductory material in this part covers a variety of subjects, including loan defaults, their risks, and their causes; machine learning's advantages and effects are briefly described; the relationship between machine learning and credit approval prediction; and strategies for reducing loan failures. The remainder of this paper will evaluate the models with the greatest accuracy ranking, emphasize the attempts made to identify the overall best model, and offer justifications for the work completed. As was previously indicated, the decision made to approve of loans has a significant influence on finance and has the potential to undermine global economic stability. Furthermore, the establishment of a stable economic environment can be facilitated by the granting of loans that are founded on accurate data and analysis [5]. Due to the vast quantities of data that expert bankers have gathered, Machine Learning is excellent at forecasting the chance that a loan will be approved. This helps the Machine Learning algorithms build the models more quickly and accurately

he following are the improvements that have been made to this topic:

- Testing five machine learning methods
- Predicting loan approval using machine learning.
- Three data-sets were used, totaling 72,215 entries with various feature sets ranging from 11 to 36 features.

The following sections of this paper are organized in the following order:first The introduction, which defines the evolving environment within the financial industry. As we move further, the second section digs into an investigation of related work, providing a full summary of previous researches. Moving on, the third section, a critical component, includes a detailed overview of the methodology used in this study, carefully explaining the data-set and the exact algorithms used for analysis. As we move to the fourth part, we reach an important point in which the study's findings and comprehensive examination of the recommended algorithms are highlighted. The fifth section summarizes the conclusion, bringing together major results and effects from the research.As we move to the sixth part expresses appreciation and acknowledgement to all the supporting personalities who have played an important role in the research's success.The seventh and final section contains all of the paper's references.

II. Related Work

The domain of loan approval prediction has been extensively studied, with numerous researchers delving into this area and achieving notable outcomes. Our research draws upon an analysis of several pertinent papers, all of which will be meticulously referenced in the citations section, to bolster and inform our investigation.

Jaymin Patel et al.[1] In this paper the importance of predicting an accurate discission of whether a loan is better to be rejected or approved is being discussed and the huge effect of loan defaults on the banking industry and economy in general . Considering the dangerous influence on finance this could lead to , authors have employed a range of machine learning approaches, including well known algorithms such

as Decision Trees, Random Forests, Gaussian Naive Bayes, Support Vector Machines, and Multiple Logistic Regression, to forecast bankruptcies. The algorithms are conducted by the research team using a mixture of metrics, including F1-score, accuracy, recall, and precision to determine the best model that could achieve the lowest error percentage. Furthermore, the team used a variety of dataset and they have reached out that he customer's employment status or years of job experience and with their debt revenue were the most effective features for forecasting loan failures. The research concluded that model prediction has a strong ability to predict bankruptcies over a large number of customers with a very high success percentage introducing a solid base for loan prediction through machine learning. To recap, the paper is a very helpful source for analyzing the effect of machine learning using a suitable algorithm on the detection of loan defaults and highlighting how this can have a positive effect on the financial state of banks by helping them reduce their loss.

Amruta S.Aphale and et al.[2]. Based on training data taken from banking system using some algorithms which is, Nearest Centroid, Gaussian Naive Bayes, Neural Networks, Discriminant Analysis, and Naive Bayes to discuss the application of machine learning for analyzing this data using previous algorithms. Finally, predicting the creditworthiness.

Archana S. and et al.[3] discussing predicting loan eligibility using machine learning and deep learning algorithms to provide the insights into effectiveness of different algorithms. Taking 70 % of training data to makes some processes using algorithms,Decision Tree (DT), Logistic Regression (LR), K Nearest Neighbour (KNN), Random Forest Classifier, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes, XGBoost, Gradient Boosting, Adaboost, Deep Neural Network (DNN), Long Short Term Memory network (LSTM). THe previous algorithms is concerned with Machine Learning (ML), and Deep Learning (DL) algorithms. And when setting testing data 30% of data to check if it is able to predict correct which means reaching the aim. After analyzing the performance of the models, found that both Logistic Regression and Linear Discriminant Analysis had the best accuracy of 82.43%.

The research paper "Loan Approval Prediction based on Machine Learning Approach" by Kumar Arun, Garg Ishan, and Kaur Sanmeet[4] from Thapar University, India. focuses on the critical need for the banking industry to have a highly accurate and efficient loan approval process. In an effort to lower risk and conserve bank resources, the authors suggest a machine learning-based method for forecasting the safety of making loans to individuals. The study provides an analysis of the performance of the six machine learning classification models, according to the authors, is the Random Forest model best predicting model. The system's possible effects on the banking industry are examined. the paper presents an accurate and supportive approach to the problems associated with predicting loan approval, which may be useful to banks as well as customer.

Diwate.[8] The author Diwate in his research paper "Loan Approval Prediction Using Machine Learning". The author started by splitting the data he brought from several banks and split it into training and testing.He used support vector machine and logistic regression algorithms to forecast the loan approval.The results concluded that support vector machine was the dominant with respect to the measurement (accuracy, precision, and recall.). Therefore,support vector machine model was outperformed by the logistic regression model in terms of results.

Suleiman Mohamed Fa-ti.[10] In the research paper "Applications of machine learning Based Prediction Model for Loan Status Approval" by the author Suleiman Mohamed Fa-ti. He used various machine learning algorithms in order to predict the loan approval. He began by Data pre-processing that contain data analysis, handling missing values getting rid of outliers.After he divided the data set into training and testing, He used machine learning algorithms as Random Forest, Decision Tree, and Logistic Regression. At the end, Logistic Regression was the dominant with respect to accuracy, precision, recall, F1 score, and Area Under the Curve .

Kavita Khadse.[13] In the research paper "Applications of machine learning in loan prediction systems" by the author Kavita Khadse used machine learning algorithms to forecast the prediction of loan.The author began by performing some data analysis and divided the data into training and testing. Moreover, she applied different machine learning algorithms. At the end Random Forest was the dominant algorithm with accuracy 94% while using the test data set and also without over fitting. Therefore, this algorithm assured as the best option for loan prediction.

Peter Nabende and Samuel Senfuma.[16]In this research paper the importance of supervised machine learning for predicting weather a loan is approved or not is explained as for the great impact of such a topic on the commercial state of the financial organizations and institutes. Authors have applied 3 methods: neural networks, Decision Tree (DT), and k-Nearest Neighbor (KNN) to discuss their efficient in bringing accurate results. These algorithms were tested on a very special data brought from a Ugandan financial institution to make their test based on a real dataset for enhancing the accuracy percentage of the models. In addition to that, the training dataset mentioned previously contained 1000 loan application records which contributed in the increasing the efficiency of forecasting loan condition. Unfortunately, the in paper didn't include which model have achieved the highest accuracy percentage for determining loan status. To summarize, the study has talked about the value of predicting a loan is approved or denied for the financial institutes. authors also used a exclusive and real dataset for their algorithms analysis to increase their precision and accuracy of their outcomes.

Golak Bihari Rath, Debasish Das , BiswaRanjan Acharya [18] The paper offers an advanced approach that banks can use to grant loans: machine learning. A dataset of previous loan applicants was gathered by the authors, who then divided it into training and testing datasets. They constructed prediction models using a variety of classification techniques, including logistic regression, decision trees, and SVM, and assessed their effectiveness using measures including accuracy, precision, recall, and F1-score. The final prediction model used was the logistic regression model since it was shown to have the highest accuracy. The authors propose that this method can offer a quick and dependable substitute for the present bank loan approval processes. The paper also includes a literature review of previous studies on loan analysis using machine learning and discusses the potential for applying this approach to other industries beyond banking.

Apurb Rajdhan et al.[21] The contribution of machine learning detecting the accuracy of loan approval percentage through implementing multiple algorithms on real customers data has a significance rule in strengthen the Customer relationship (CRM).selecting the suitable algorithm is the core point in the process ; for guarantying the lowest errors percentage. As a result, in this paper different algorithms are being compared to determine the perfect one in terms f accuracy. Furthermore, SVM has ranked the highest algorithm in accuracy percentage over others that were: KNN, NN, and DT. In a nutshell, the paper shows the importance of machine Learning in the improvement of loan approval prediction process in banks and reducing their loss using data mining techniques based on real data by choosing the most proper algorithm .

Devanch Shah et al.[23] In this research Abhishek Kumar Tiwari's is highlighting the effect of advanced machine learning and statistics over the achievement in the loan improvement process in banks through the facilitation of degerming the ability of a customer to repay back or not. The authors related to the Great Recession that happened on year 2008 ,were banks were in critical state as a result of loan defaults and this lead to a great break down in the economy. Moreover, to limit the amount of bankruptcies loans must be given through a trusted discission .Furthermore, Here comes the rule of machine learning where it brings over a trustworthy and an easy discission for loan approval applying a great progress in banks . The progress that happened was done through analyzing the efficient of multiple algorithms on massive datasets for real customers from banks that were supposed to huge amount of loss and training the model on them to choose the most effective and accurate one for such a job .To boot four methods were being tested and compared in this investigation ; which are: , the research uses four machine learning approaches: Random Forest (RF), K-Nearest Neighbors (KNN), Classification and Regression Tree (CART), and Logistic Regression decision tree classifiers and the importance processing data and cleaning them before testing was being discussed a. The results showed that the Random Forest and KNN models especially have achieved the highest accuracy giving the merit of the success to machine learning in this field . In addition, the study didn't only suggest the best model but it listed the advantages and disadvantages of each model along with there default percentage. The results of the four algorithms were being compared according to four matrices : (AUC), F1 score, Recall, Precision, and Accuracy as for the huge importance for achieving the highest precise results as for the sensitivity of the topic and it's great impact on economy. To sum up, the study is very valuable as it suggests a precise analysis and comparison to multiple algorithms used in machine learning to asset banks and support them in the loan approval matter while choosing the suitable and trusted customer for accepting his loan.

Nazim Uddin, Md. Khabir Uddin Ahamed, Md Ashraf Uddin, Md. Manwarul Islam, Md. Alamin Talukder, Sunil Aryal[24] The paper meant to present banking loan approval using intelligent application-based on machine learning. The paper discusses the difficulties banks have in determining which loan applicants are worthy and offers a new approach that makes use of an interface and ensemble learning. Nine machine learning algorithms and three deep learning models are used in the study. the data-set is balanced and model performance is enhanced. The accuracy, precision, recall, and f1_score of the system are assessed. financial institution and clients can easily and swiftly checks the loan status. According to the study, the suggested system offers a practical way to boost the loan approval procedure and increase the accuracy of bank loan approval forecasts

A research is conducted in [9] to determine which clients will and won't repay their loans. Loan repayment behavior is analyzed and predicted using a variety of machine learning methods, including Random Forest, Decision Tree, and Logistic Regression. With an evaluation accuracy of 89.7059%, Logistic Regression was shown to perform the best, followed by Decision Tree (85.4054%) and Random Forest (77.4566%). The study also identifies the variables—such as loan size and credit history—that affect loan repayment. According to this paper, the best machine learning approach for loan prediction is logistic regression. The study shows how machine learning may be used to forecast loan payback behavior and offers insightful information for the lending sector.

The use of machine learning models to predict loan approval for banking companies is also covered in the paper [22]. Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), Linear Models (LM), Neural Networks (Nnet), and Adaboost (ADB) are the six machine learning classification models presented in this study. The test dataset consists of newly submitted application details, while the training dataset is utilized to

train the machine learning model. The models are assessed according to their capacity to forecast the likelihood that a new applicant will be approved for a loan. The ability to get the money back is a fit case. Loan ID, Gender, Married, Dependents, Education, Applicant Income, Coapplicant Income, Loan Amount, Loan Amount_Term, Credit History, Property_Area, and Loan_Status are among the variables in the dataset.Additionally, the article offers code examples for data analysis and manipulation using the Pandas module in Python. It also talks about how difficult it is for banking institutions to anticipate credit defaulters and approve loans. The accuracy, precision, and recall metrics obtained from the six algorithms' experiments are shown, and the conclusion emphasizes how effective the predictive models are. According to the report, the application can be incorporated into other systems and satisfies the needs of banking institutions.The study reveals that, when compared to the previously described methodologies, the Decision Tree (DT) machine learning algorithm had the highest accuracy. Therefore, the Decision Tree model was found to be the most successful in forecasting loan approval for new applicants based on the experimental data reported in the research. It also talks about how the prediction module might be improved in the future. All things considered, the study gives a thorough review of machine learning methods for predicting loan acceptance and sheds light on the difficulties and possibilities in this field.

This paper [15] presents a Loan Approval Prediction System using Machine Learning, which aims to classify safe customers for loan approval using some algorithm. The system uses two datasets; a training dataset to build a model and a testing dataset to provide accurate results for loan approval. The document discusses the use of the algorithm in the Loan Approval Prediction System. It highlights that the system uses the algorithm, which builds a model from the training dataset and applies it to the testing dataset to obtain the required output. The algorithm is describing ensemble approach and building a strong model . The paper also claims the discussion of limitations of existing loan approval techniques and the benefits of using machine learning algorithms in the banking institutions. The system is secure and reliable to be used in banking situations and can be applied in the real world

III. PROPOSED METHODOLOGY

Numerous algorithms were used, and a research was done on each algorithm before training the model using them on the data-sets. The following diagram represents the steps the data-sets went through to get the results.

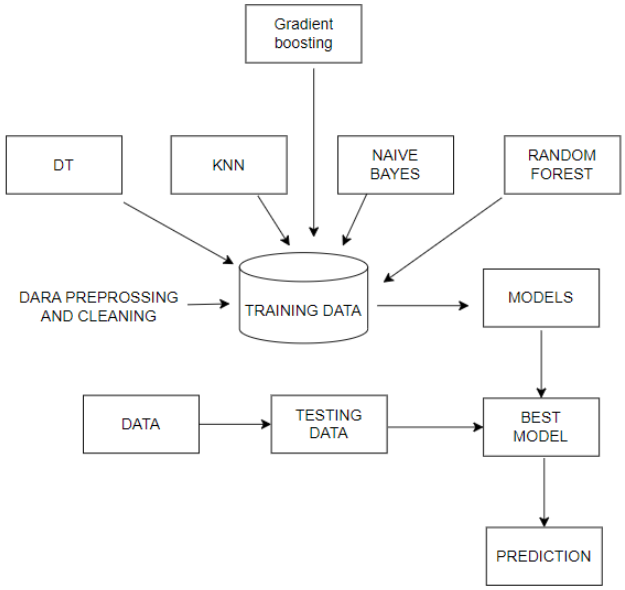

Fig. 4. Loan Approval prediction

A. *Data-sets Descriptions*

The first data-set named "loan approval prediction" has 12 features and was divided into two partitions: 80% for training and 20% for testing. A thorough summary of the characteristics follows.

These characteristics give critical information for forecasting loan acceptance. The number of dependents, education level, self-employment status, income, loan amount, loan term, credit score, and asset valuations are all variables that might affect loan acceptance. Researchers can uncover patterns and linkages that contribute to the accuracy of the prediction model and determine which criteria are important in deciding loan acceptance by evaluating these characteristics. The goal variable that signals the loan approval outcome is the loan status feature, and it is critical for training the prediction model.

TABLE I
FEATURES OF LOAN APPROVAL PREDICTION DATASET

Features	Type	Values
loan_id	Numerical	From 1 to 4269
no_of_dependents	Numerical	Interval of numbers(0,1,2,3,4,5)
education	Classification	Graduate or Not Graduate
self_employed	Classification	Yes or No
income_annum	Numerical	From 200,000 to 9,900,000
loan_amount	Numerical	From 300,000 to 4E+07
loan_term	Numerical	Interval of numbers(2,...,18,20)
cibil_score	Numerical	From 300 to 900
residential_assets_value	Numerical	From -100,000 to 2.9E+07
commercial_assets_value	Numerical	From 0 to 1.9E+07
luxury_assets_value	Numerical	From 300,000 to 3.9E+07
loan_status Classification Approved or Rejected		

The second data-set named "Loan Default dataset" has 15 features. The data-set was standardized before being divided into two parts: 80% for training and 20% for testing. A thorough summary of the characteristics follows.

ID: A unique identification assigned to each loan application. Loan Amount: The loan amount sought. Funded Amount: The loan's actual funding amount. Investor Funded Amount: The amount contributed by investors. Term: The length of the loan. Batch Enrolled: The loan's batch identity. The interest rate that has been applied to the loan. The loan has been awarded a grade. Sub Grade: The loan's designated subgrade. Employment length: The applicant's employment length. The sort of house ownership. Verification Status: The status of the applicant's information's verification. Payment Plan: Indicates whether or not the loan has a payment plan. Loan Title: The title that has been issued to the loan. Debit to Income: This is the debt-to-income ratio.

TABLE II
FEATURES OF LOAN DEFAULT PREDICTION DATASET

Features	Type	Values
ID	Numerical	Identifier
Loan Amount	Numerical	From 1014 to 35,000
Funded Amount	Numerical	From 1014 to 34,999
Funded Amount Investor	Numerical	From 1114.59 to 34,999.75
Term	Classification	36,58,59
Batch Enrolled	Numerical	Identifier
Interest Rate	Numerical	From 5.320006 to 27.18235
Grade	Classification	A, B, C, D, E,F
Sub Grade	Classification	A[1:4],B[1:4],C[1:4],D[1:4],F[1:4]
Employment Duration	Classification	Mortgage or Rent or Own
Home Ownership	Numerical	Interval of numbers(39833.921,..)
Verification Status	Classification	Source Verified,Verified or Not Verified
Payment Plan	Classification	n
Loan Title	Classification	Credit card refinancing or others
Debit to Income	Numerical	Interval of numbers(7.914332762,..)
Delinquency - two years	Numerical	Interval of numbers(0,..,6)
Inquires - six months	Numerical	Interval of numbers(0,..,6)
Open Account	Numerical	Interval of numbers(0,..30)
Public Record	Classification	0 or 1
Initial List Status	Classification	f or w
Loan Status	Classification	0 or 1

The third dataset named "Loan Approval " with an excel file name : cleaned dataset has 12 features.The dataset was standardized before being divided into two parts: 80% for training and 20% for testing. A thorough summary of the characteristics follows.

Loan_ID: A unique identification assigned to each loan application. Gender: The applicant's gender. Married: Indicates whether or not the candidate is married. Dependents: The applicant's number of dependents. Education: The applicant's level of education. Self_Employed: Indicates whether or not the applicant is self-employed. ApplicantIncome: The applicant's income. CoapplicantIncome: The co-applicant's income. LoanAmount: The loan amount sought. Loan_Amount_Term: The loan's duration. Credit_History: The applicant's credit history. Property_Area: The neighborhood in which the property is located. Loan Status: Indicates whether or not the loan was authorized.

TABLE III
FEATURES OF CLEANED_DATA DATASET

Features	Type	Values
loan_id	Numerical	Identifier
no_of_dependents	Numerical	Interval From 1 to 5
education	Classification	Graduate or Not Graduate
self_employed	Classification	Yes or No
income_annum	Numerical	From 200,000 to 9,900,000
loan_amount	Numerical	From 300,000 to 39,500,000
loan_term	Numerical	From 2 to 20
cibil_score	Numerical	From 300 to 900
residential_assets_value	Numerical	from _-100,000 to 29.100,000
Numerical commercial_assets_value	Numerical	From 0 to 19,400,000
luxury_assets_value	Numerical	From 300,000 to 39,200,000
bank_asset_value	Numerical	From 0 to 14,700,000
loan_status	Classification	Approved or Rejected

B. *Used Algorithms*

The aforementioned datasets were fed into six distinct machine learning algorithms: Naive Bayes, Random Forest, K Nearest Neighbor (k-NN), Gradient Boosting, Random Forest, and Decision Tree. The following statistics were produced for every algorithm: Accuracy, Recall, Precision, and f1_score. The outcomes were then compared and recorded. The paper's results, graphics, and discussion of the results are located later on.The aforementioned datasets were fed into six distinct machine learning algorithms: Naive Bayes, Random Forest, K Nearest Neighbor (k-NN), Gradient Boosting, Random Forest, and Decision Tree. The following statistics were produced for every algorithm: Accuracy, Recall, Precision, and f1_score. The outcomes were subsequently contrasted and recorded.The analysis and investigation will be found in the upcoming sections[14]

1) **Gradient Boosting:**
Gradient boosting is a type of boosting utilized in machine learning. It is built on the presumption that when the head potential future model is matched with former versions, the overall prediction error is lessened. To minimize error, the rudimentary concept is to define the target results for the succeeding model. [7]. Also an overview about it:First step is The initial setup: For every sample in the data set, the algorithm makes an initial prediction. For regression, this first prediction might be the target variable's mean; for classification, it can be the class with the highest frequency.Second step is Creating Weak Models: The process iterative and constructs a number of weak models, most frequently decision trees. Every weak model aims to fix the errors committed by previous versions.Third step is Residual computation: For every sample in the data set, the algorithm computes the residuals—the difference between the expected and actual values—in each iteration. The mistakes or declassification's committed by the current ensemble of weak models are represented by these residuals.Fourth step is Fitting Weak Models to Residuals: Next, the residuals are used to fit the weak models. The aim is to develop a model that

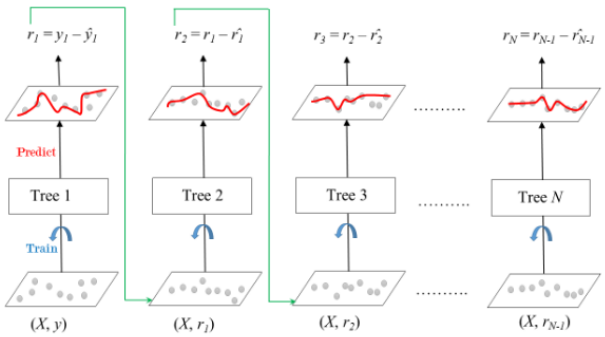

Fig. 5. Gradient Boosting

outperforms the prior ensemble of weak models in terms of residual prediction accuracy.Fifth step is Weak Model Weighting: Based on how well a weak model predicts the residuals, it is awarded a weight. A weak model's weight in the final ensemble increases with its ability to forecast the residuals.Sixth step is Upgrading Ensemble Forecasts: The ensemble forecast is updated by combining the predictions made by each weak model. The contribution of each weak model to the final prediction is controlled by multiplying the predictions made by the current weak model by a learning rate.Seventh step: is Iterative Process: Until a stopping requirement is satisfied, steps 3-6 are repeated a predetermined number of times. By including new weak models in the ensemble, the algorithm attempts to lower the residuals and enhance the overall forecast in each iteration.Final Prediction: The Gradient Boosting algorithm's final prediction is derived by adding the predictions of each of the ensemble's weak models.

Advantages :

- High Predictive Accuracy
- Handles Complex Relationships.
- Handles Missing Data.
- Flexible

Disadvantages :

- Sequential not parallel
- It has Over fitting risk.
- Expensive in computation.
- Time consuming

2) **Decision Tree:**
The Decision Tree algorithm is a sophisticated and adaptable supervised machine learning approach that may be used for classification and regression tasks. For illustrating decision paths, this algorithm relies on a hierarchical, tree-like architecture. In this complicated network, each core node represents an important decision based on a specific feature, while each leaf node represents the final conclusion or forecast.

The Decision Tree's core lies on its iterative approach for data division. The algorithm cautiously identifies the most significant feature within the data set at each stage and strategically separates the data into subgroups. This recursive process is repeated until specified requirements are fulfilled, leading in the development of a tree structure that professionally captures the data's complicated patterns and connections. As a result, the Decision Tree algorithm emerges as a sophisticated tool that provides a complete and understandable. approach to data-driven decision-making in the field of machine learning.

Classification tree: When the projected outcome is the class to which the data belongs, an analysis is performed. For example, the outcome of a loan application may be classified as safe or dangerous.

Regression tree: When the anticipated outcome is an actual number, the analysis is complete. Consider the population of a state.

Advantages:

- Decision trees need less work for data preparation during pre-processing than other methods.
- Data normalization is not required for a decision tree.
- A decision tree does not need data scalability.
- Missing values in the data have no significant impact on the process of developing a decision tree.
- The Decision Tree Model is highly straightforward and simple to communicate to both technical teams and stakeholders.

Disadvantage:

- A minor change in the data might result in a significant change in the structure of the decision tree, resulting in instability.
- When compared to other algorithms, the computation for a decision tree might be significantly more difficult at times.
- The training period for a decision tree is frequently longer.
- Decision tree training is highly costly due to the increased complexity and time required.
- The Decision Tree technique is insufficient for predicting continuous values and performing regression.

[12].

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \left| \frac{|S|}{S|} \cdot \text{Entropy}\,(S_v) \right. \tag{1}$$

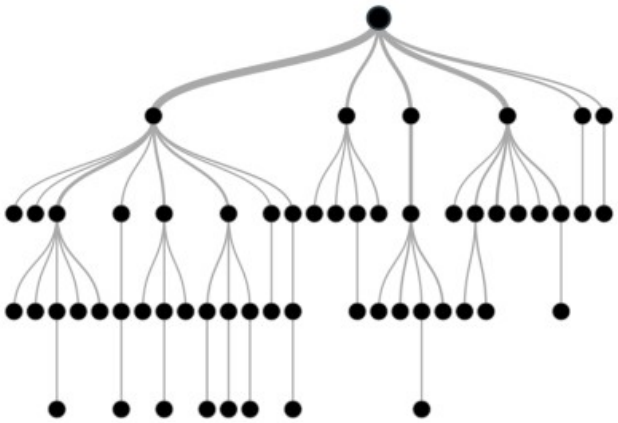

Fig. 6. Illustration of decision tree[11]

3) **Naive Bayes:**
Built on the Bayes Theorem, Naive Bayes is a simple yet capable categorization algorithm, which describes the probability of an event based on prior knowledge of conditions that might be related to the event. It presupposes predictor independence, which means that the traits or characteristics are not related to one another or connected in any way. Even though there is a dependence, each of these qualities or attributes contributes to the probability independently, which is why it is termed Naive. [6]

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \tag{2}$$

$$P(X,Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)c) \tag{3}$$

$$P(D \mid A,S) = \frac{P(A \mid D,S)P(D \mid S)}{P(A \mid S)} \tag{4}$$

Advantages and Disadvantages of Naive Bayes:

1-Advantages:

-Efficient with small data sets.
-Fast training and prediction.
-Handles irrelevant features.
-Works well with categorical features.
-Can handle missing data.

2-Disadvantages:

-Zero probability issue.
-Limited expressive power.
-Sensitive to irrelevant features.
-Requires a large amount of data for some variations.

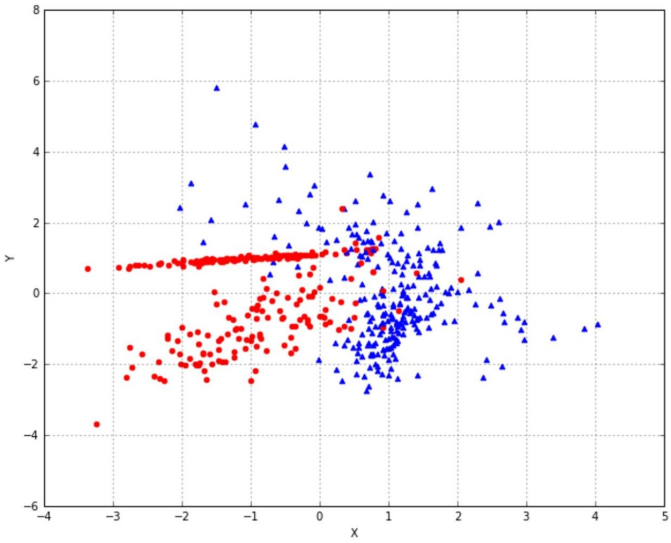

Fig. 7. Naive Bayes[20]

4) **K – Nearest Neighbor:**
Among of the simplest yet most important categorization methods in machine learning is K-Nearest Neighbors. It is heavily used in intrusion detection , pattern recognition, and data mining, and represents a member of the supervised learning field. K-NN fails to make any assumptions about the underlying data since it is a non-parametric method. Because it saves the data set and acts on it while classifying, it is also known as a lazy learner algorithm. This is because it doesn't absorb information from the data it was trained on right away.[25].For classification and regression problems, a wellliked machine learning method is the Kliked-Nearest Neighbor (K-NN) algorithm. It is predicated on the notion that comparable points of data typically have comparable labels or values.
The K-NN method uses the complete training data sets as a reference throughout the learning phase. It uses a selected distance metric, such as Euclidean distance, Manhattan Distance, and Minkowski Distance to determine the distance between each training sample and the data being used as input points before generating recommendations.
Considering their distances, the algorithm then determines the K closest neighbors of the given data point. When it comes to classification, the algorithm predicts the label for the input data point based on the most prevalent class label among the K neighbors.In order to forecast the value for the data being used as an input point in regression, the average or weighted average of the target values of the K neighbors is computed.
Euclidean Distance:

$$\text{distance}(x, X_i) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{5}$$

Manhattan Distance

$$\text{distance}(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{6}$$

Minkowski Distance:

$$\text{distance}(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{7}$$

[17]

5) **Random Forest:**
Within supervised learning, the random forest classifier has become a strong and adaptable method that is highly valued for its ability to perform both regression and classification tasks. Using a random feature selection and a distinct portion of the training data for each decision tree, this technique makes use of the collective knowledge of the group. Because of this diversity, there is less chance of over-fitting, which is a drawback of single decision trees and results in better generalized and performance on unknown data. The ensemble character of the random forest is its essential premise. Instead of building a single decision tree, it builds numerous separate ones, each using a random subset of features for splitting at each node and a bootstrapped sample of the original data. By introducing a random element, the trees become more diverse and are unable to over optimize on the unique training data that each one of them receives. As a result, the final prediction is the result of adding together each tree's individual predictions, either by averaging for regression or a majority vote for classification. The collective wisdom surpasses the constraints of individual trees, resulting in forecasts that exhibit a measurable increase in accuracy and resilience against noise. The random forest classifier's attractiveness goes beyond its remarkable accuracy in predictions. Its exceptional flexibility, which easily accommodates both numerical and category features within its fold, is its fundamental strength. Furthermore, its skill in handling high-dimensional data makes it an excellent choice for working with contemporary datasets that are feature-rich. Moreover, the random forest structure is easily interpreted and provides insightful information on feature relevance metrics. This capability enables practitioners and academics to better understand the underlying model and hone their analysis by revealing the features that have the biggest impact on the predictions. Even though random forest classifiers have clear benefits, it's important to recognize their drawbacks. The success of these models can only be maximized by meticulous hyperparameter tuning, which can be a costly and time-consuming procedure computationally. Furthermore, training big tree ensembles can be computationally demanding, which could present problems in environments with limited resources. In conclusion, the random forest classifier offers a compelling blend of accuracy, robustness, and adaptability, making it a cornerstone technique in the machine learning landscape. Its ability to excel in both classification and regression tasks, coupled with its inherent resistance to overfitting and its amenability to high-dimensional data, positions it as a formidable tool for modern data analysis. While mindful consideration of its hyperparameter tuning requirements and computational demands is essential, the random forest classifier undoubtedly deserves its prominent place within the arsenal of data-driven methodologies.

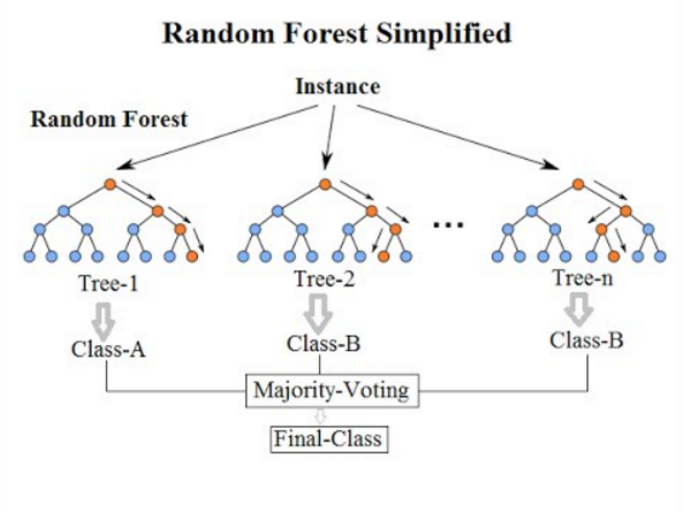

Fig. 8. Random forest demonstration [19]

C. *Performance Metrics*

Accuracy is the count of legitimately anticipated data from all the data. The count of accurately anticipated positives taken from the anticipated positives is the Precision Recall is the number of correctly anticipated positives from all the true positives. The number of accurately anticipated negatives out of all the expected negatives is known as specificity.

$$\text{Accuracy} = (\text{TN}+\text{TP})/(\text{TN}+\text{TP}+\text{FN}+\text{FP}) \tag{8}$$

$$\text{Precision} = \text{TP}/\text{TP}+\text{FP} \tag{9}$$

$$\text{Recall} = \text{TP}/\text{TP}+\text{FN} \tag{10}$$

IV. Results and Analysis

The results collected from the Decision tree, K-nearest neighbor, Random Forest, Gradient Boosting, and Naive Bayes are shown below.

TABLE IV
STATISTICS OF ALGORITHMS FOR LOAN APPROVAL PREDICTION DATASET

Model	Accuracy	Precision	Recall	f1-score
Decision Tree	0.9778	0.9777	0.9778	0.9754
K-NN	0.5574	0.5574	0.5574	0.5574
Random Forest	0.9766	0.9766	0.9766	0.9766
Gradient Boosting	0.9906	0.9906	0.9906	0.9906
Naive Bayes	0.7681	0.7681	0.7681	0.7681

The following results are from the first dataset.

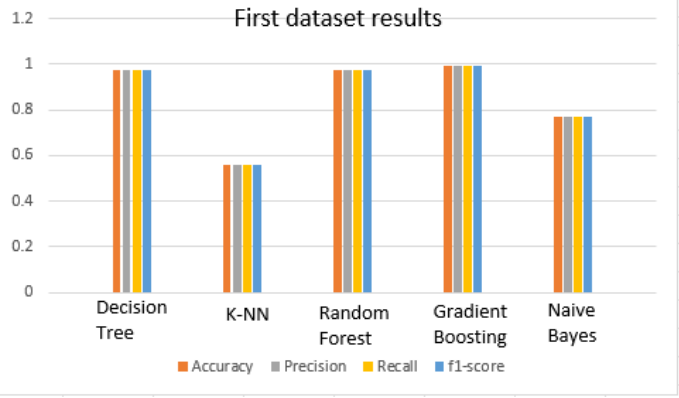

Fig. 9. Loan Approval Prediction dataset performance chart with data split

Gradient Boosting(GB), Decision Tree, and Random Forest(RF) are the dominant algorithms, they all have accuracy greater than 95%. Gradient Boosting is the best approach for this data set with an accuracy of 99%. The Decision Tree and the Random Forest are the second best approaches for this data set with an accuracy of 98%. The Naive Bayes is not the best approach but it is also not bad it can work with an accuracy of 77%. The K-NN is the worst approach for this data set with an accuracy of 56%. The precision, the recall, and the f1-score are almost equal to the accuracy in the five approaches.

The following results are from the Second dataset.

TABLE V
STATISTICS OF ALGORITHMS FOR THE LOAN DEFAULT PREDICTION DATASET

Model	Accuracy	Precision	Recall	f1-score
Decision Tree	0.82598	0.8367	0.82598	0.8312
K-NN	0.9034	0.1161	0.0107	0.0195
Random Forest	0.9065	0.9065	0.9065	0.9065
Gradient Boosting	0.8021	0.8025	0.9559	0.8725
Naive Bayes	0.9005	0.9005	0.9005	0.9005

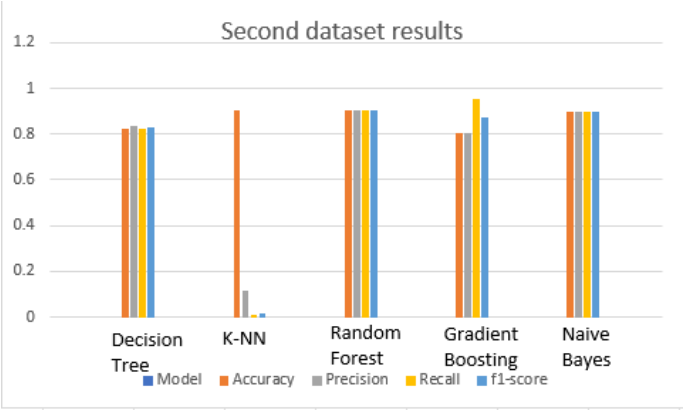

Fig. 10. Loan Default Prediction dataset performance chart with 10 k-fold

Random Forest(RF), Naive Bayes are the dominant algorithms, they all have accuracy greater than 90%.Random Forest(RF) is the best approach for this data set with accuracy 90%. The Naive Bayes is the second best approaches for this data set with accuracy 90%.The K-NN is the worst approach for this data set with accuracy 1.95%. The precision, the recall and the f1-score is almost equal to the accuracy in the five approaches.

The following results are from the cleaned data dataset.

Model	Accuracy	Precision	Recall	f1-score
Decision Tree	0.75	0.74	0.75	0.75
K-NN	0.625	0.7	0.8235	0.7568
Random Forest	0.7396	0.7396	0.7396	0.7396
Gradient Boosting	0.8021	0.8021	0.8021	0.8021
Naive Bayes	0.9896	0.9896	0.9896	0.9896

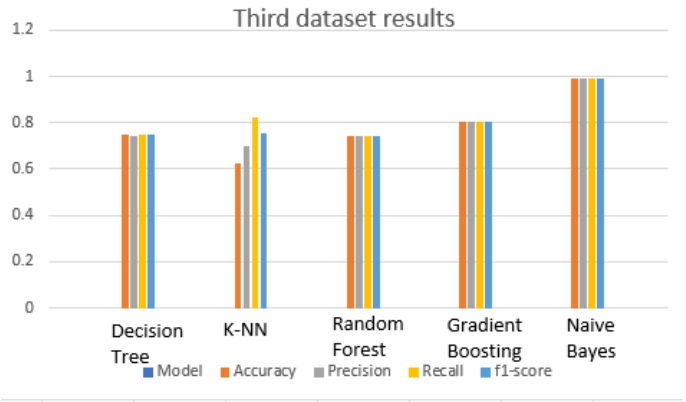

Fig. 11. cleaned data data-set performance chart with 10 k-fold

Naive Bayes is the dominant algorithms, It has accuracy 99%.Decision tree is the second best approach with accuracy 75%.Then the Random Forest has accuracy 73%. And the worst approach is the K-NN with 63%. The best approach for the whole data set is the Naive Bayes.

V. Conclusion

"In conclusion, it is impossible to overestimate the importance of financial decision-making, especially when it comes to predicting loan acceptance. This study used three different data sets to develop a unique machine learning framework for loan prediction. We conducted a comparison research between the Gradient Boosting Classifier and Decision Tree, Naive Bayes, K-NN , Random Forest, and other algorithms. The results showed that the Gradient Boosting Classifier method performed better than the others every time. The most appropriate option for dependable and accurate loan acceptance forecasts is the Gradient Boosting Classifier, which has an accuracy of 90% for the first and second data sets and 80% for the third.

VI. Acknowledgment

First and foremost, we would like to show our gratitude to the staff working at Misr International University and the faculty of computer science for working hard to make this university a success. We are very thankful to Prof. Mohamed Shebl ElKomy University President, Prof. Ayman Nabil dean of faculty of Computer Science, and Prof. Abdelnasser Zaied Vice Dean of Student Affairs and Professor of Computer Engineering for helping us to learn in this virtuous university for running it flawlessly. And finally, we would like to give special thanks to Dr. Diaa AbdelMoneim associate professor in information systems, Eng. Mohamed Khaled, and Eng. Tarek Mohamed teaching assistants for teaching us such valuable information for their continued guidance and support and for our work.

References

[1] Mayank Anand, Arun Velu, and Pawan Whig. Prediction of loan behaviour with machine learning models for secure banking. *Journal of Computer Science and Engineering (JCSE)*, 3(1):1–13, 2022.

[2] Amruta S Aphale and Sandeep R Shinde. Predict loan approval in banking system machine learning approach for cooperative banks loan approval. *International Journal of Engineering Trends and Applications (IJETA)*, 9(8), 2020.

[3] S Archana and KS Divyalakshmi. A comparison of various machine learning algorithms and deep learning algorithms for prediction of loan eligibility.

[4] Kumar Arun, Garg Ishan, and Kaur Sanmeet. Loan approval prediction based on machine learning approach. *IOSR J. Comput. Eng*, 18(3):18–21, 2016.

[5] Maria Balgova, Michel Nies, and Alexander Plekhanov. The economic impact of reducing non-performing loans. 2016.

[6] Thomas Bayes. Naive bayes classifier. *Article Sources and Contributors*, pages 1–9, 1968.

[7] Navoneel Chakrabarty, Tuhin Kundu, Sudipta Dandapat, Apurba Sarkar, and Dipak Kumar Kole. Flight arrival delay prediction using gradient boosting classifier. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2*, pages 651–659. Springer, 2019.

[8] Yash Diwate, Prashant Rana, and Pratik Chavan. Loan approval prediction using machine learning. *International Research Journal of Engineering and Technology (IRJET)*, 8(05), 2021.

[9] Prateek Dutta. A study on machine learning algorithm for enhancement of loan prediction. *International Research Journal of Modernization in Engineering Technology and Science*, 3, 2021.

[10] Suliman Mohamed Fati. Machine learning-based prediction model for loan status approval. *Journal of Hunan University Natural Sciences*, 48(10), 2021.

[11] Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In *icml*, volume 99, pages 124–133, 1999.

[12] Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora, and Naresh Dhami. Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications*, 163(8):15–19, 2017.

[13] Kavita Khadse. Applications of machine learning in loan prediction systems. *Linguistica Antverpiensia*, 3:3658–3674, 2020.

[14] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386, 2020.

[15] PL Srinivasa Murthy, G Soma Shekar, P Rohith, and G Vishnu Vardhan Reddy. Loan approval prediction system using machine learning. *Journal of Innovation in Information Technology*, 4(1):21–24, 2020.

[16] Peter Nabende and Samuel Senfuma. A study of machine learning models for predicting loan status from ugandan loan applications. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pages 462–468. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2019.

[17] Swathi Nayak, Manisha Bhat, N V Subba Reddy, and B Ashwath Rao. Study of distance metrics on k - nearest neighbor algorithm for star categorization. *Journal of Physics: Conference Series*, 2161(1):012004, jan 2022.

[18] Golak Bihari Rath, Debasish Das, and BiswaRanjan Acharya. Modern approach for loan sanctioning in banks using machine learning. In *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019*, pages 179–188. Springer, 2021.

[19] Steven J Rigatti. Random forest. *Journal of Insurance Medicine*, 47(1):31–39, 2017.

[20] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[21] DURGESH KUMAR SINGH and NOOPUR GOEL. Customer relationship management: Two dataset comparison in perspective of bank loan approval using machine learning techniques. *Journal of Theoretical and Applied*

Information Technology, 101(19), 2023.

[22] J Tejaswini, T Mohana Kavya, R Devi Naga Ramya, P Sai Triveni, and Venkata Rao Maddumala. Accurate loan approval prediction based on machine learning approach. *Journal of Engineering Science*, 11(4):523–532, 2020.

[23] Abhishek Kumar Tiwari. Machine learning application in loan default prediction. *JournalNX*, 4(05):1–5, 2018.

[24] Nazim Uddin, Md Khabir Uddin Ahamed, Md Ashraf Uddin, Md Manwarul Islam, Md Alamin Talukder, and Sunil Aryal. An ensemble machine learning based bank loan approval predictions system with a smart application. *International Journal of Cognitive Computing in Engineering*, 4:327–339, 2023.

[25] Shichao Zhang, Debo Cheng, Zhenyun Deng, Ming Zong, and Xuelian Deng. A novel knn algorithm with data-driven k parameter computation. *Pattern Recognition Letters*, 109:44–54, 2018.