# DefBDet: An Intelligent Default Borrowers Detection Model

Fooz Alghamdi[1], Nora Alkhamees[2]
Zakat, Tax and Customs Authority, Riyadh, Saudi Arabia[1]
Department of Management Information Systems, King Saud University, Riyadh, Saudi Arabia[2]

***Abstract*—The growing popularity and availability of online lending platforms have attracted more borrowers and lenders. There have been several studies focusing on analyzing loan risks in the financial industry, however, defaulting loans still remains an issue that needs more attention. Hence, this research aims to develop an intelligent prediction model that is able to predict risky loans and default borrowers, named the Default Borrowers Detection Model (DefBDet). We seek to help loan lending platforms to approve lending loans to those who are expected to comply with re-payments at the agreed time. Previous works developed a binary classification prediction model (either default or repaid loan). Repaid loans include loans being repaid on or after the loan deadline date. DefBDet, on the other hand, is a novel model, it can predict a loan status based on a multi-classification bases rather than a binary class bases. Hence, it can additionally identify expected late repaid loans, so that special conditions are assigned before loan being approved. This study employs seven different Machine Learning models, using a real-world dataset from 2009-2022 consisting of around 255k loan requests. Statistical measures such as Recall, Precision, and F-measure have been used for models' evaluation. Results show that Random Forest has achieved the highest performance of 85%.**

***Keywords*—*Default borrowers; default loans; loan risks; machine learning models; prediction model***

## I. INTRODUCTION

Nowadays, data more than ever before is being generated at an extremely fast rate, even more; the volume of data being produced every day is extraordinary [1] [2]. Over the next few years up to 2025 [3], global data creation is expected to grow to more than 180 zettabytes [3]. Furthermore, the global financial data analytics market was valued at $7.6 billion in 2020 and is forecasted to reach $19.8 billion (more than the double value) by 2023 [4].

Default loans data-based assessment is being widely used in financial organizations around the world to assist organizations in either approving or rejecting loan requests [5]. In addition, the growing popularity of loans shows that, in the US [6], more than 20 million persons were owed $178 billion in personal loans as of the first quarter of 2022. That's more than the double of what was in 2015, only $88 billion were owed from personal loans [6].

One of the reasons for the rising demand for lending is the availability and accessibility of online lending platforms, also the simplicity of completing the loan applications process (with no or limited conditions), which led to a jump in loan requests in 2022 [6]. Lending is risky in that repayments are not always guaranteed, thus, increasing the number of defaulters which is expected to reach 6% in 2023 [7][8]. It has been noticed that the number of willful default cases has increased in 2021 [9], compared with the earlier two years. In addition, loans are considered the dominant asset in the banking sector, they represent nearly 75% of the total amount of the banking assets [10].

Therefore, it is indeed critical to manage the lending activities in a way that controls the borrower's compliance and maintains the financial institution's performance, assets, and liquidity. In contrast, failure to manage loan compliance would likely affect the economy at large [10].

These reasons push toward the need to detect defaulters at early stages (i.e. prior to the loan request being approved), which essentially means identifying the loan requester who will be potentially a defaulting borrower. This would definitely help loan lending organizations to approve loan requests of only committed borrowers and conserve financial institutions' resources. The above-mentioned challenges confirm the importance of monitoring loan requests by developing an intelligent default borrower's identification model with an early warning system that is capable of alerting financial institutions of potential losses and preventing crises, which in turn is the aim of this paper.

Machine Learning (ML) plays an essential role in the financial sector, especially in the loans domain [11]–[21]. In this work, we develop an intelligent prediction model that is able to recognize the compliance of borrowers against the loan's repayment. Also, it is able to identify expected late repaid loans (loan being fully repaid after loan's maturity date). Thus, we develop a novel model named, Default Borrowers Detection Model (DefBDet). This model is able to predict the loan status based on a multi-classification bases rather than binary class bases, which was the case in previous works [11]–[21].

Hence, the main objective, in this work, is to improve the utilization of classification algorithms in the financial loans field. We seek to assess default loan risk prior to issuing the loan lending decision. ML techniques have been employed in order to develop DefBDet using real loan data. We used the family of supervised ML algorithms, such as Decision Tree (DT), Random Forest (RF), ID3, Deep Learning (DL), Gradient Boosted Decision Trees (GBDT), Support Vector Machines (SVMs), and Naïve Bayes (NB).

The rest of the paper is organized as follows. The next section provides a summary of past related studies. Later, the processes involved in building the model, including data collection and processing, model training and testing, and associated findings and results are presented. The final section concludes with a summary of the work and the suggested plan for future work.

## II. Literature Review

ML has been utilized intensively in the financial industry along with its subfields, including but not limited to, the stock market [22]–[24], insurance [25]–[27], and fraud detection [28]–[34].

The stock market field is an attractive topic due to the abundance of data being generated at high and irregular rates. Besides, it plays an important role in the economy; investors are also continually looking to predict future transactions to avoid certain associated risks. The study in [24] has constructed a model to predict stock market future trends. In addition, along with historical stock market prices it considered sentiment analysis using text polarity of financial news. Regarding the insurance field, this study [27] has summarized the role of Data Mining (DM) in the insurance industry and how DM enhances the decision-making using insurance data. For fraud detection, it can be committed in different ways and areas, such as in banking, insurance, government, and healthcare sectors [34]. The paper in [28] provides a comprehensive review of existing research works and literatures on the applications of DM to fraud detection in finance.

### A. *Predicting Default Loans and Risk Detection*

Businesses use predictive analytics to identify potential risks and opportunities for their organizations [35]. Recently, it has been observed that the number of willful defaulters in the financial sector is significantly rising [36]. Therefore, systematic identification to predict and detect willful default behavior is indeed essential. The Corporate Finance Institute [37] has defined the loan as: is the sum of money that one or more individuals or companies borrow from banks or other financial institutions, to financially manage planned or unplanned events. In doing so, the borrower must pay back with interest and within a specified period of time [37]. Following we show some research performed in this domain, this including but not limited to [11]–[21].

Due to the increase in using the Internet and submitting loan applications online, it is defiant for financial institutions to evaluate all loans manually. Thus, predicting whether a borrower is going to default is becoming an extremely urgent need and draws much attention from researchers. The author in [11] proposed a supervised default loan prediction method based on deep metric learning, the method extracts the features of a loan itself, models the hidden relationships in loan pairs, and calculates the probability of default. The challenges were the imbalanced defaulted samples compared with total data of loans, hard decision boundaries due to loans binary label, and the heterogeneous loan features that have different data types. The proposed method has a higher accuracy compared to Support Vector Machines, Logistic Regression, Naive Bayes, and Multi-layer Perceptron.

Another work in [12], the aim was to identify a comprehensive list of factors along with building a data model for early prediction of whether the loan will become a Non-Performing Assets (NPA)[1] or not. They explored different classification techniques and considered Neural Network, because of its higher accuracy. The model covers the loan end-to-end processes, starting from loan request where the factors of NPA are early detected, followed by loan monitoring, where the model can identify outliers and possible requests to be defaulted. And the last stage is closing the loan, whether it is fully repaid or declared as NPA.

In [13], the authors presented a study for predicting whether a peer-to-peer (P2P) loan application will be repaid or defaulted by employing different classification models. This work was aiming to find interesting relations among the attributes of loan application by applying association rules. The used dataset by LendingClub[2] was classified as whether a loan will default or not (Yes/No). The most effective classification model was achieved using Random Forest and its accuracy was 71.75%.

Using the LendingClub dataset, a two-phase model is proposed in [14], the first phase predicts loan acceptance or rejection by applying Logistic Regression with recall score of 77.4%. The second phase, on the other hand, predicts loans requests either will be defaulted or fully paid by applying Deep Neural Network with recall score of 72%.

Another study [15] wants to predict loan defaulters using the LendingClub dataset. The author encodes loan status (Current, Fully paid, Issued) as Normal, and encodes (Default, Charged off, In Grace Period, Late) as Default. The performance evaluation was applied and compared using Random Forest algorithm with other three ML methods, namely, Decision Tree, Logistic Regression, and Support Vector Machines, while Random Forest still performs the best.

In another prediction work on LendingClub dataset [16], researchers label the dataset records as follows: any loan that (Defaulted, Charged off, or Late on repayments) was classified as Negative examples, while classified any loan that was (Fully paid or Current) as Positive examples. Naïve Bayes has been used which performs the best with default prediction rate compared with other models.

This study [17] presented clustering for loan risk analysis on big data using the k-mean clustering algorithms. Researchers used different datasets related to loans, including the Bondora[3] dataset. The clustering of Bondora was divided into two classes: Default risk and Non-default risk.

[1] A non-performing asset (NPA): is a classification used by financial institutions for loans and advances on which the principal is past due and on which no interest payments have been made for a period of time. – Corporate Finance Institute.

[2] LendingClub is a peer-to-peer lending company headquartered in San Francisco, California.

[3] Bondora is one of the leading non-bank digital consumer loan providers in Europe.

The author in [18] developed a DM model for predicting loan default among P2P loans. The work was specifically for small business owners and employed using Boosted Decision Tree model. In this study, the class labels are (Pay in full and Charged off).

Another work utilized Bondora dataset [19] is mainly focused on the prediction of loan default using ML algorithms. The author used Status attribute to predict loan default, which is an existing attribute in the dataset. Status reflects the loan payment status and has polynomial values as follows: Current, Repaid, and Late. The work converted the polynomial values into binary, while the loan records that having Repaid status are treated as Not default and Late status are treated as Default, whereas the records having Current status are excluded as they play no role in the default classification.

In view of this and after reviewing the above literature, the majority of studies [11]–[21] have classified the loans into binary classification (i.e. repaid or not), also, they have issues in determining the loan payment activities and the default stages. In more detail, the Repaid status in previous works does not only consider that the loan was paid on the agreed time (i.e. loan repaid before the loan deadline), but also considers late loan payment (i.e. payment after the loan deadline, such as in the case of default) as Repaid too.

Furthermore, it is worth noting that Repaid status does not reflect exactly whether or not the borrower defaulted prior to the loan being fully repaid. In this case, the default behavior is not detected properly. Alternatively, additional investigation is required in the dataset to determine the actual loan status whether default behavior happened or not before the loan is flagged as Repaid, as it is the key challenge to predicting the defaulters. There is a crystal-clear necessity for multi-classifications, such as Late Repaid classification, while it is not applicable that we include the Late Repaid records in Repaid classification. As a result, having a precise multi-classification will improve the quality of the loan prediction model and can help financial institutions to accurately determine the compliance stage of the borrower, and therefore, take the decision to lend or not. Multi-classification of repayment can differentiate between lenders who are compliant to close their loans on the agreed time and lenders who are delayed repaying the loan.

## III. Data Description

### A. Data Collection and Data Description

Bondora is known to be one of the leading non-bank digital consumer loan providers in Europe and has been operating since 2009 [38]. It is licensed as a credit provider under the Estonian Financial Supervision Authority [39].

It is worth noting that Bondora dataset has been extensively used in different published works, including but not limited to [17], [19], [40]–[43].

Bondora's real-world data is publicly available on the internet[4] [44]. It contains raw data related to loan requests from 2009 till today (download day is 22/9/2022) and it is daily updated. The dataset consists of 122 attributes and around 255k records.

Each record belongs to a loan request, which includes data about the borrower and the loan application, such as borrower demographical data, loan duration, purpose of the loan, dates of payment, and also information regarding the current loan's status (Repaid, Late, Current), amount of principal and interest, default date, etc.

Due to the enormity of the dataset, we analyzed it in-depth and determined the most relevant attributes to the study's purpose. Table I provides a full description of the selected attributes in the dataset.

TABLE I. Full Description of Bondora's Attributes in the Dataset [44]

| Feature Name | Brief Description | Data Type |
|---|---|---|
| LoanId | A unique ID given to all loan applications | Categorical |
| PartyId | A unique ID given to the borrower | Categorical |
| NewCreditCustomer | Did the customer have prior credit history in Bondora<br>False: Customer had at least 3 months of credit history in Bondora<br>True: No prior credit history in Bondora | Categorical |
| LoanDate | Date when the loan was issued | Date |
| MaturityDate_Original | Loan maturity date according to the original loan schedule | Date |
| MaturityDate_Last | Loan maturity date according to the current payment schedule | Date |
| Age | The age of the borrower when signing the loan application | Numeric |
| Gender | Male, Female, and Undefined | Categorical |
| Amount | Amount the borrower received on the Primary Market. This is the principal balance of your purchase from Secondary Market | Numeric |
| LoanDuration | Current loan duration in months | Numeric |
| UseOfLoan | Loan consolidation, Real estate, Home improvement, Business, Education, Travel, Vehicle, Health, and Other | Categorical |
| Education | Primary education, Basic education, Vocational education, Secondary education, and Higher education | Categorical |
| NrOfDependants | Number of children or other dependents | Categorical |
| EmploymentStatus | Unemployed, Partially employed, Fully employed, Self-employed, Entrepreneur, and Retiree | Categorical |
| OccupationArea | Other, Mining, Processing, Energy, Utilities, Construction, Retail and wholesale, Transport and warehousing, Hospitality and catering, Info and telecom, Finance and insurance, Real-estate, Research, Administrative, Civil service & military, Education, Healthcare and social help, Art and entertainment, Agriculture, and Forestry and fishing | Categorical |
| HomeOwnershipType | Homeless, Owner, Living with | Categorical |

[4] To download the loan dataset (https://www.bondora.com/en/public-reports#secondary-market-archive)

| | | |
|---|---|---|
| | parents, Tenant_pre-furnished property, Tenant_unfurnished property, Council house, Joint tenant, Joint ownership, Mortgage, Owner with encumbrance, and Other | |
| IncomeTotal | Borrower's total income in (€) | Numeric |
| LastPaymentOn | The date of the current last payment received from the borrower | Date |
| DefaultDate | The date when the loan went into defaulted state and collection process was started.<br>Or, in other words, the loan is 60+ days overdue | Date |
| RecoveryStage | Current stage according to the recovery model<br>1 Collection<br>2 Recovery<br>3 Write Off | Categorical |
| Status | The current status of the loan application<br>Repaid: the loan is fully paid to the investor<br>Current: the loan is still in the process<br>Late: the loan has not been fully paid on due dates | Categorical |

While examining the dataset, we discovered that 53 records had missing values in the majority of the attributes.

Due to the limited number of rows, we decided to exclude them from the dataset. The removed records are approximately less than 0.25% of the total dataset.

Also, some attributes in the dataset were numeric data type, and to facilitate the analysis, we converted these attributes from numeric to categorical data type by following discretization process [45]. Table II shows each attribute with original values before applying discretization and the transformed values after the discretization process.

TABLE II. BONDORA'S ATTRIBUTES AFTER DISCRETIZATION PROCESS

| Attribute | Previous Values | New Values | Corresponded Values |
|---|---|---|---|
| Age [46] | Values range from 18 to 77 | Young adults Middle-aged adults Old-aged adults Seniors | (From age 18 to 28) (29 to 39) (40 to 58) (>=59) |
| IncomeTotal | Values range from 0 to 1,012,019 | No income Very low income Low income Middle income High income | (Total between 0 to 10) (11 to 1,000) (1,001 to 2,000) (2,001 to 3,000) (>= 3,001) |
| Amount | Values range from 6 to 15,948 | 0 - 1,500 1,501 - 3,000 3,001 - 4,500 4,501 - 6,000 More than 6,000 | |
| LoanDuration | Values range from 1 to 120 | 1 – 30 31 – 60 61 – 90 91 - 120 | |
| NrOfDependants | Values range from 0 to More than 10 | No dependent One – Three Four – Six Seven – Ten More than Ten Undefined | |
| NrOfPreviousLoans | Values range from 0 to 74 | No previous loan One – Five Six – Ten Eleven – Fifteen Sixteen – Twenty More than Twenty Undefined | |

Bondora's dataset provides Loan Status, which indicates the loan payment status, it has three possible values, Repaid (meaning that loan has been fully paid back to the investor either on or before the loan maturity date, or after the loan maturity date), Late (the loan has not yet been fully paid to the investor and has exceeded the loan maturity date), and Current (the loan is still in progress, i.e. maturity date has not been reached yet). Almost 30% of the dataset consists of Late loans, and 35% of both Current loans and Repaid loans, Fig. 1 illustrates in a pie chart the ratio of each status of the total dataset.

The original dataset lacks the ability to show the borrowers status regarding paying the loan on time or not. Hence, if the Loan Status is Repaid, it does not clearly indicate whether the loan was repaid on time (on or before the loan maturity date) or not. In other words, it does not show the cooperation of the borrower to repay the loan by the maturity date. Knowing that the loan was repaid on time saves the lending company resources and means that the borrower is trustworthy and dependable.

As a result, and since the dataset has no direct indicator regarding the borrower’s compliance in repaying the given loan on the agreed time, we defined a new polynomial feature named "Borrower's compliance status", which precisely represents the situation of the loan in terms of repayment according to the data in each record.

The new feature measures the borrower compliance regarding paying the loan's last payment before the maturity date of the loan or not.

It consists of four values, Repaid on-time, Late repaid, Defaulted, and in progress.

A loan can be labelled as "Repaid on-time", meaning that the loan was repaid before or on the maturity date of the loan (the dataset consist of 76k record or almost 30% of records were labelled as Repaid on-time), "Late repaid" on the other hand, which means that the loan was repaid after the loan maturity date (15k or 6% of records were labelled as Late repaid). Also, a loan can be labelled as "Defaulted", meaning that the loan is not fully repaid yet, and has passed the maturity date of the loan too (71k or 28% of records were labelled as Defaulted), or labelled as "In progress", which means the current loan is still open and not meet any of the above status (93k or 36% of records were labelled as In progress).

This feature was defined using the following existing attributes in the dataset: MaturityDate_Original, LastPaymentOn, DefaultDate, and Status.

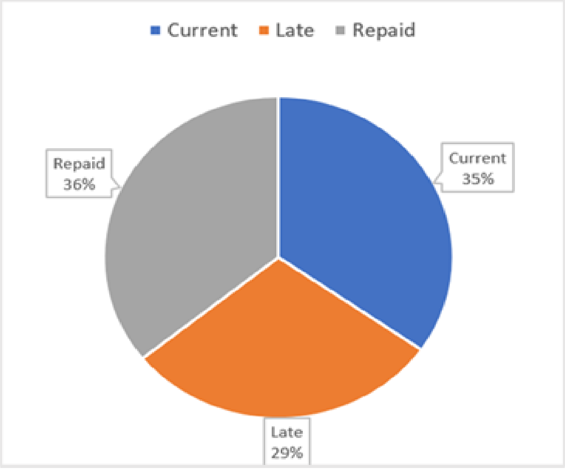

Fig. 1. Proportions of loan status of bondora's dataset.

## IV. Experiment and Methods

### A. Machine Learning Classification Algorithm

ML was developed by Arthur Lee Samuel in 1959 [47], the author demonstrated that machines could learn from past errors.

In this study, ML algorithm was determined after reviewing existing works and literatures in this scope [11]–[21], in addition to the study's purpose which is to predict the borrowers' compliance against the loan re-payment.

Developing a prediction model is indeed valuable for financial institutes and banks as it enables business owners to take actionable decisions at the right time. It helps to avoid borrowers who have common characteristics with other defaulters. Thus, preventing tangible and intangible losses that are expected to occur, or in critical conditions, escalating the case to the court. Therefore, the prediction model will support and enhance the approval or the rejection of loan applications that are expected to default.

In this research, we applied various classification algorithms to build DefBDet which belongs to the family of supervised ML algorithms. These algorithms are DT [48]–[50], ID3 [51], [52], RF [53]–[55], DL, GBDT [18], [56], SVMs [57], [58], and NB [59].

### B. Model Evaluation and Assessment

In this study, we mainly focus on several performance evaluation metrics for classifier including F-measure, Precision, Recall. Below are brief definitions of each measure:

- According to [60], F-measure was first introduced by Cornelis Joost van Rijsbergen [61], it combines Recall and Precision with an equal weight [62]. F-measure formula as following (1) [60]:

$$F1 = (2 * Recall * Precision) / (Recall + Precision) \quad (1)$$

- Precision or Confidence is the proportion of predicted positive cases that are correctly real positives [62]. Precision formula as following (2) [63]:

$$Precision = (\sum TP) / (\sum TP + FP) \quad (2)$$

Recall or Sensitivity is the proportion of real positive cases that are correctly predicted positive. This measures the Coverage of the real positive cases [62]. Recall formula as following (3) [63]:

$$Recall = (\sum TP) / (\sum TP + FN) \quad (3)$$

Accuracy is a common measure for evaluating a classification model's ability to discriminate between classes. And because the class labels are imbalanced, the F- measure will capture a balance between Recall and Precision and weights them equally better than the Accuracy in terms of describing the overall model performance [64], [65]. The evaluation metrics and performance measures of each algorithm are detailed in Table III.

## V. Results and Discussion

In this study, we have employed ML classification technique (a supervised learning method) in order to predict the compliance status of loan borrowers. We used seven different classification algorithms and evaluated its performance by F-measure, Precision, and Recall.

The results show that the RF and DT have outperformed all other classification algorithms. In addition, they both have very comparable results (the obtained results for seven algorithms are displayed in Table III). The DT has the highest F-measure score of 85.69%, while RF ranked in the second place with very minor differences of 85.29%.

It is worth noting that the outcomes of all performance measures are considered essential and crucial. However, from business prescriptive, specifically the financial field, we are concerned that the loan is provided to the client who will compliance and is willing to repay on time. The financial organizations do not want to lose this type of client (borrowers), i.e. financial organizations are always keen to provide loans to clients who comply with payments on time. Thus, in terms of a performance measure, this is described as precision, which predicts positive cases that are correctly real positives [62]. In other words, the model can predict only the positive borrowers (who will repay on time), than identifying all positive borrowers.

Financial organizations are not willing to provide a loan to someone who has a slight probability of defaulting, and this is due to the importance of managing the financial, human, and digital resources effectively to be sustained in the market, besides the organization's reputation among competitors.

In DefBDet, the precision based on RF is above 92%, on the other hand, the precision of DT is almost 89%, indicating that the RF model has a strong ability for detecting correct positive cases than DT.

Other reasons why we choose RF algorithm over the other ML algorithms are that RF runs efficiently on enormous datasets and works well with unbalanced class labels. We can conclude obviously that the RF algorithm outperforms other ML approaches according to the mentioned reasons above.

TABLE III. EVALUATION METRICS COMPARISON OF THE SEVEN ALGORITHMS

| Rank based on F-measure | Algorithm | F-measure | Precision | Recall |
|---|---|---|---|---|
| 1 | DT | 85.69% | 88.93% | 82.67% |
| 2 | RF | 85.29% | 92.14% | 79.39% |
| 3 | GBDT | 84.19% | 88.93% | 79.92% |
| 4 | NB | 79.52% | 77.09% | 82.11% |
| 5 | DL | 77.95% | 73.24% | 83.30% |
| 6 | ID3 | 77.22% | 83.54% | 71.78% |
| 7 | SVMs | 59.92% | 54.78% | 66.13% |

## VI. CONCLUSION AND FUTURE WORK

This study has developed an intelligent prediction model identifying default loans in lending communities; its main aim was improving the lending decision-making process. In other words, DefBDet model aims to detect the expected default borrower before the approval of the loan request, which can reduce default loans and at the same time maintain the expected target of returns.

We can clearly infer that DefBDet could reduce the defaulting loans with positive consequences for the efficiency of the financial institutions, by eliminating loan requests that have been detected and are expected to default. Furthermore, it can identify loans that are expected to be late repaid (i.e. loans being repaid after the loan maturity date has passed). Late repaid loans can be an issue to financial institutions if not being closely monitored.

Previous works were able to binary classify loan requests to either fully repaid or defaulted. The case of fully repaid, includes both loans repaid on or before the loan deadline and loans being repaid after the loan deadline date has passed. However, we think expected Late repaid loans needs special attention before loan approval being issued. Thus, DefBDet is a multi-class model; it aims to identify expected late repaid borrowers, so that additional conditions and/ or close monitoring are given. DefBDet can classify a loan to Repaid on-time, Late repaid, and Default.

In addition, the results provided by the model can be generalized to any lending activities or financial institutions. In general, the DT and RF algorithms showed a better performance, the overall performance was higher than 85%, compared with other classification models like SVMs, ID3, GBDT, NB, and DL. However, from a business point of view, correctly identifying defaulting borrowers is very crucial; it can lead to saving financial institutions resources. Financial organizations are not willing to provide a loan to someone who has a probability of defaulting. Thus, high precision value is indeed favorable. In DefBDet, the precision using RF was above 92%, on the other hand, the precision of DT was nearly 89%. As a result, The RF algorithm was adopted in DefBDet.

In the future, we seek to employ DefBDet prediction model on a local dataset to explore the diversities between international and local datasets. In addition, we aim to introduce an additional borrower class (label), which leads to improving DefBDet to be able to predict an additional multiclass label in order to precisely identify the compliance stage of the borrowers before repayment.

## ACKNOWLEDGMENT

The authors would like to thank the Researchers Supporting Project number (RSP2023R212), King Saud University, Riyadh, Saudi Arabia.

## REFERENCES

[1] “Data Mining: What it is and why it matters | SAS.” https://www.sas.com/en_sa/insights/analytics/data-mining.html (accessed May 18, 2022).

[2] B. Vuleta, “How Much Data Is Created Every Day? +27 Staggering Stats,” 2021. https://seedscientific.com/how-much-data-is-created-every-day (accessed Jan. 09, 2023).

[3] “Total data volume worldwide 2010-2025 | Statista,” 2022. https://www.statista.com/statistics/871513/worldwide-data-created/ (accessed Jun. 06, 2022).

[4] K. Kanhaiya, R. Pradeep, and K. Vineet, “Financial Analytics Market Size| Industry Forecast - 2030,” 2022. https://www.alliedmarketresearch.com/financial-analytics-market (accessed Dec. 26, 2022).

[5] V. Singh, A. Yadav, R. Awasthi, and G. N. Partheeban, “Prediction of Modernized Loan Approval System Based on Machine Learning Approach,” 2021 International Conference on Intelligent Technologies, CONIT 2021, Jun. 2021, doi: 10.1109/CONIT51480.2021.9498475.

[6] M. SCHULZ, “Personal Loan Statistics: 2022 | LendingTree,” 2023. https://www.lendingtree.com/personal/personal-loans-statistics/ (accessed Dec. 26, 2022).

[7] “2023 U.S. Lev Loan Default Forecast Raised to 2.0%-3.0%; 2024 Projected at 3.0%-4.0%,” Fitch Ratings, 2022. https://www.fitchratings.com/research/corporate-finance/2023-us-lev-loan-default-forecast-raised-to-2-0-3-0-2024-projected-at-3-0-4-0-30-09-2022/ (accessed Jan. 13, 2023).

[8] “Default, Transition, and Recovery: The U.S. Speculative-Grade Corporate Default Rate Could Reach 3.75% By September 2023 | S&P Global Ratings.” https://www.spglobal.com/ratings/en/research/articles/221121-default-transition-and-recovery-the-u-s-speculative-grade-corporate-default-rate-could-reach-3-75-by-sept-12565939 (accessed Jan. 13, 2023).

[9] “Wilful default cases down by over 50% in last eight years: Govt data - Times of India,” 2021. https://timesofindia.indiatimes.com/business/india-business/wilful-default-cases-down-by-over-50-in-last-eight-years-govt-data/articleshow/90409785.cms (accessed Dec. 26, 2022).

[10] A. Abaidoo and S. Oppong, Determinant of Loan Default and Its Effect on Financial Performance of Commercial Banks in Ghana. A Case Study of Fidelity Bank Limited. 2017.

[11] K. Zhuang, S. Wu, and X. Gao, “A deep metric learning approach for weakly supervised loan default prediction,” Journal of Intelligent and Fuzzy Systems, vol. 41, no. 4, pp. 5007–5019, 2021, doi: 10.3233/JIFS-189987.

[12] G. Attigeri, M. M. Manohara Pai, and R. M. Pai, “Framework to predict NPA/Willful defaults in corporate loans: A big data approach,” International Journal of Electrical and Computer Engineering, vol. 9, no. 5, pp. 3786–3797, Oct. 2019, doi: 10.11591/ijece.v9i5.pp3786-3797.

[13] Z. Alomari, “Loan Default Prediction and Identification of Interesting Relations between Attributes of Peer-to-Peer Loan Applications,” vol.2, 2017.

[14] J. D. Turiel and T. Aste, “P2P Loan acceptance and default prediction with Artificial Intelligence,” Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.01800

[15] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, “A study on predicting loan default based on the random forest algorithm,” in Procedia Computer Science, Elsevier B.V., 2019, pp. 503–513. doi: 10.1016/j.procs.2019.12.017.

[16] S. Chang, S. Dae-Oong Kim, and G. Kondo, “Predicting Default Risk of Lending Club Loans,” 2015.

[17] K. Kumar Pandey and D. Shukla, “STRATIFIED REMAINDER LINEAR SYSTEMATIC SAMPLING BASED CLUSTERING

MODEL FOR LOAN RISK DETECTION IN BIG DATA MINING," International Journal of System Assurance Engineering and Management, vol. 13, 2021, doi: 10.1007/s13198-021-01424-0.

[18] A. Semiu and A. A. R. Gilal, "A boosted decision tree model for predicting loan default in P2P lending communities," Int J Eng Adv Technol, vol. 9, no. 1, pp. 1257–1261, Oct. 2019, doi: 10.35940/ijeat.A9626.109119.

[19] V. Padimi, .. V. S., and D. D. Ningombam, "Applying Machine Learning Techniques To Maximize The Performance of Loan Default Prediction," Journal of Neutrosophic and Fuzzy Systems, pp. 44–56, 2022, doi: 10.54216/JNFS.020204.

[20] A. Jafar Hamid and T. M. Ahmed, "Developing Prediction Model of Loan Risk in Banks Using Data Mining," Machine Learning and Applications: An International Journal, vol. 3, no. 1, pp. 1–9, Mar. 2016, doi: 10.5121/mlaij.2016.3101.

[21] S. Samsir, S. Suparno, and M. Giatman, "Predicting the loan risk towards new customer applying data mining using nearest neighbor algorithm," in IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, May 2020. doi: 10.1088/1757-899X/830/3/032004.

[22] N. Alkhamees and M. Aloud, "Intelligent Algorithmic Trading Strategy Using Reinforcement Learning and Directional Change," 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9514595 (accessed Dec. 02, 2022).

[23] N. Alkhamees and M. Aloud, "DCRL: Approach for Pattern Recognition in Price Time Series using Directional Change and Reinforcement Learning," (IJACSA) International Journal of Advanced Computer Science and Applications, 2021, Accessed: Dec. 02, 2022. [Online]. Available: www.ijacsa.thesai.org

[24] A. E. Khedr, S. E. Salama, and N. Yaseen, "Predicting stock market behavior using data mining technique and news sentiment analysis," International Journal of Intelligent Systems and Applications, vol. 9, no. 7, pp. 22–30, Jul. 2017, doi: 10.5815/ijisa.2017.07.03.

[25] Bharat. Rao, Balaji. Krishnapuram, A. (Andrew) Tomkins, Q. Yang, and Association for Computing Machinery. Special Interest Group on Knowledge Discovery & Data Mining., Data Mining to Predict and Prevent Errors in Health Insurance Claims Processing. Association for Computing Machinery, 2010.

[26] K. P. M. L. P. Weerasinghe and M. C. Wijegunasekara, "A Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims," European International Journal of Science and Technology, vol. 5, no. 1, 2016, Accessed: Nov. 06, 2022. [Online]. Available: www.eijst.org.uk

[27] K. Umamaheswari and S. Janakiraman, "Role of Data mining in Insurance Industry," 2014.

[28] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," Decis Support Syst, vol. 50, no. 3, pp. 559–569, 2011, doi: 10.1016/j.dss.2010.08.006.

[29] S. Patil and R. Bhowmik, "Detecting Auto Insurance Fraud by Data Mining Techniques," vol. 2, no. 4, 2011, [Online]. Available: http://www.cisjournal.org

[30] N. Carneiro, G. Figueira, and M. Costa, "A data mining based system for credit-card fraud detection in e-tail," Decis Support Syst, vol. 95, pp. 91–101, Mar. 2017, doi: 10.1016/j.dss.2017.01.002.

[31] L. Mialaret et al., "Neural data mining for credit card fraud detection," 1999.

[32] G. L. Gray and R. S. Debreceny, "A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits," International Journal of Accounting Information Systems, vol. 15, no. 4, pp. 357–380, 2014, doi: 10.1016/j.accinf.2014.05.006.

[33] P. K. Chan and S. J. Stolfo, "Toward Scalable Learning with Non-uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection," 1998. [Online]. Available: www.aaai.org

[34] A. Gillis, "What is fraud detection? Definition from SearchSecurity," 2021. https://www.techtarget.com/searchsecurity/definition/fraud-detection (accessed May 20, 2022).

[35] "What Is Predictive Analytics? | Alteryx." https://www.alteryx.com/glossary/predictive-analytics (accessed May 20, 2022).

[36] ENS Economic Bureau, "Wilful defaulters rise by over 200 to 2,494 in FY21: Nirmala Sitharaman | Business News,The Indian Express," 2021. https://indianexpress.com/article/business/banking-and-finance/wilful-defaulters-rise-by-over-200-to-2494-in-fy21-nirmala-sitharaman-7425706/ (accessed May 20, 2022).

[37] CFI Team, "Loan - Definition, Types and Things to Consider Before Applying," 2023. https://corporatefinanceinstitute.com/resources/commercial-lending/loan/ (accessed Dec. 02, 2022).

[38] "What is the general business information of Bondora? | Bondora Support." https://support.bondora.com/en/what-is-the-general-business-information-of-bondora (accessed May 20, 2022).

[39] "Bondora is now regulated by the Estonian FSA." https://www.bondora.com/blog/bondora-is-now-regulated-by-the-estonian-financial-supervision-authority/ (accessed May 31, 2022).

[40] Š. Lyócsa, P. Vašaničová, B. Hadji Misheva, and M. D. Vateha, "Default or profit scoring credit systems? Evidence from European and US peer-to-peer lending markets," Financial Innovation, vol. 8, no. 1, Dec. 2022, doi: 10.1186/s40854-022-00338-5.

[41] J. Mezei, A. Byanjankar, and M. Heikkilä, Credit risk evaluation in peer-to-peer lending with linguistic data transformation and supervised learning. 2018. [Online]. Available: http://hdl.handle.net/10125/50056

[42] A. Byanjankar and M. Viljanen, "Predicting expected profit in ongoing peer-to-peer loans with survival analysis-based profit scoring," in Smart Innovation, Systems and Technologies, Springer Science and Business Media Deutschland GmbH, 2019, pp. 15–26. doi: 10.1007/978-981-13-8311-3_2.

[43] I. Czarnowski, R. J. Howlett, and L. C. Jain, "Smart Innovation, Systems and Technologies 193 Intelligent Decision Technologies Proceedings of the 12th KES International Conference on Intelligent Decision Technologies (KES-IDT 2020)," 2022. [Online]. Available: http://www.springer.com/series/8767

[44] Bondora, "Public Reports | Bondora." https://www.bondora.com/en/public-reports (accessed Apr. 14, 2023).

[45] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features," ICML, vol. 1995, 1997, doi: 10.1016/B978-1-55860-377-6.50032-3.

[46] "Age Categories, Life Cycle Groupings," 2017. https://www.statcan.gc.ca/en/concepts/definitions/age2 (accessed May 20, 2022).

[47] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," IBM Journal of, vol. 3, no. 3, pp. 210-229, July 1959, doi: 10.1147/rd.33.0210.

[48] S. K. Murthy, "Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey," Data Min. Knowl. Disc.. vol. 2, no. 4, 2000, doi: 10.1023/A:1009744630224.

[49] S. R. Safavian and D. Landgrebe, "A Survey of Decision Tree Classifier Methodology," IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991, doi: 10.1109/21.97458. 1990.

[50] J. R. Quinlan, "Learning Decision Tree Classifiers," ACM Computing Surveys, vol. 28, pp. 71-72, 1996, doi: 10.1145/234313.234346.

[51] J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, pp. 81-106, 1986, doi: 10.1007/BF00116251.

[52] H. Chen, G. Shankaranarayanan, L. She, and A. Iyer, "A Machine Learning Approach to Inductive Query by Examples: An Experiment Using Relevance Feedback, ID3, Genetic Algorithms, and Simulated Annealing" Journal of the American Society for Information Science, vol. 49, 1998.

[53] T. Ho, "Random decision forests," Proceedings of the 3rd International Conference on Document Analysis and Recognition, 14-16 August 1995.

[54] L. Breiman, "Random Forests," Machine Learning, vol. 45, 2001, doi: 10.1023/A:1010933404324.

[55] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, May 27, 2015. doi: 10.1038/nature14539.

[56] S. Si, H. Zhang, S. S. Keerthi, D. Mahajan, I. S. Dhillon, and C.-J. Hsieh, “Gradient Boosted Decision Trees for High Dimensional Sparse Output,” 2017. [Online]. Available: https://github.com/Microsoft/LightGBM

[57] B. Boser, I. Guyon, and V. Vapnik, “A training algorithm for optimal margin classifiers” Proceedings of the fifth annual workshop on Computational learning theory, pp. 144 – 152, July 1992, doi: 10.1145/130385.130401.

[58] R. Batuwita and V. Palade, “CLASS IMBALANCE LEARNING METHODS FOR SUPPORT VECTOR MACHINES,” Imbalanced Learning: Foundations, Algorithms, and Applications , IEEE, 2013, pp.83-99, doi: 10.1002/9781118646106.ch5.

[59] S. Taheri and M. Mammadov, “Learning the naive bayes classifier with optimization models,” International Journal of Applied Mathematics and Computer Science, vol. 23, no. 4, pp. 787–795, 2013, doi: 10.2478/amcs-2013-0059.

[60] J. D. M. Rennie, “Derivation of the F-Measure,” 2004. [Online]. Available: http://mathworld.wolfram.com/HarmonicMean.html

[61] C. J. Van Rijsbergen, “INFORMATION RETRIEVAL,” Butterworth-Heinemann, 1979.

[62] D. M. W. Powers and Ailab, “EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION,” vol. 2, no. 1, pp. 37–63, 2011, [Online]. Available: http://www.bioinfo.in/contents.php?id=51

[63] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognit Lett, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[64] J. Brownlee, “How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification,” 2020. https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/ (accessed May 21, 2022).

[65] E. Bloedorn, “Learning Rules from Highly Unbalanced Data Sets Related papers,” Fourth IEEE International Conference on Data Mining (ICDM'04), Brighton, UK, 2004, pp. 571-574, doi: 10.1109/ICDM.2004.10015.