IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"Hybrid Deep Learning Model for Credit Default Prediction Using Behavioral, Transactional, and Demographic Data"

Samiksha Katkar, Santosh Gaikwad
Department of Computer Application
JSPM University, Pune, India

Abstract:

Default prediction has become an essential task when it comes to the risk management of financial institutions and decision-making within the banking industry. In this day and age, due to advancements in technology such as digital banking services, there has been an explosion of structured and unstructured data being produced. One issue regarding machine learning algorithms is their inability to manage the complex relationships that can be found within data. Therefore, this paper suggests a hybrid deep learning algorithm which utilizes behavioral, transactional, and demographic data to predict credit default rates. This paper introduces a combined learning method that uses both Artificial Neural Networks and ensemble learning to boost the precision of predictions.

Keywords: Credit Defaults, Deep Learning, Hybrid Model, ANN.

Under such conditions, local Microgrids can be set up to meet the electricity demand. In order to cater to the demands of such remote places, multiple Microgrids are needed. Under such conditions, the creation of a smart grid, i.e., connecting several Microgrids with a battery management system, can help solve many problems related to the electricity supply. In addition, such a system takes into account the utilization of different sources of renewable energy for electricity generation in various places. The suggested method uses a long short-term memory (LSTM) algorithm for the ANN so that there is a constant electricity supply of good quality at all load buses. This paper also suggests an artificial intelligence-operating system to help manage energy under different situations. In order to improve the voltage quality, a seven-level aligned multilevel inverter has been introduced. The suggested EMS model has been validated through hardware-in-the-loop simulation using OPAL-RT modules.

I. **Introduction**:

Credit risk analysis is an essential process in the current financial world where there has been a considerable increase in the number of financial transactions using digital platforms. It is becoming increasingly challenging to predict the likelihood that a loan will be defaulted upon by a customer. Statistics cannot be used to determine hidden patterns in large volumes of data. As such, advanced models such as Deep Learning and Hybrid Models are being utilized to achieve more accurate results. A hybrid model will be employed in the project for accurate results while analyzing the likelihood of a loan being defaulted upon by borrowers. Credit risk analysis is an important process in the modern world where financial stability and profitability have become crucial. Credit risk defaults occur when a borrower fails to repay a loan. This situation leads to financial losses for the lender. Firstly, statistical models like Logistic Regression assume very few things and are not capable of dealing with any non-linear relationship.

On the other hand, machine learning models can improve prediction accuracy; however, they have limitations when it comes to handling complex data types. While deep learning models are capable of managing more intricate relationships, they need optimization and integration with other types of models. This study seeks to develop a hybrid model that is based on deep learning and involves more than one source of financial data to enhance the prediction accuracy.

In the modern era, credit risk assessment is an important component of banks and financial institutions. The advancements made by technology,

leading to digital banking, online payments, and lending services resulted in a significant increase in customer data, both structured data like income and loan history and unstructured data or semi-structured data behavior-related.

Logistic Regression is the approach that has already been utilized when forecasting credit risks. However, it is based on the presumption of linear relations between dependent and independent variables and cannot analyze complicated associations between them. In connection with the multidimensional nature of the financial data available, it is important to develop a new approach that would allow analyzing them adequately.

Using Machine Learning algorithms to solve tasks associated with regression made it possible to analyze nonlinear data and large amounts of data. Nevertheless, there can be certain limitations connected with using some machine learning approaches, such as decision trees or random forests, in case of analyzing complicated data sets.

Today credit risk assessment is an essential component of banks' activities. With the advancement of technologies, the emergence of electronic banking, online banking, and online loans led to the accumulation of great amounts of customer data, including structured (e.g., income level, credit history) and unstructured (e.g., behavior). Logistic Regression approach was used for predicting the probability of customer defaulting. However, since it presupposes linear relationships between factors, its application can lead to poor results obtained. That is why there is a need to develop other means for analyzing the data available.

II. Literature Review:

Credit default prediction is one of the most important research topics in the area of financial analytics. Logistic Regression is traditionally applied in previous literature due to its interpretability. However, Logistic Regression may suffer from nonlinearity issues in modeling nonlinear relationships.

Decision Trees and Random Forest models are applied to enhance the accuracy of the model. The problem of overfitting may be faced in applying Decision Trees and Random Forest models. SVM classifiers are used for credit default predictions in the field of financial analytics. SVM demonstrates excellent performance in handling high dimensional features.

The development of deep learning models, including ANN and LSTM models, shows excellent performance in the field of financial analytics. Hybrid models combining both deep learning and ensemble methods have been applied. Compared with traditional models, these models demonstrate superior predictive performance. The current study is based on the concept of applying hybrid models, including data from various sources.

Prediction of defaulters in credit scoring is one of the fields that have been widely analyzed by scientists dealing with financial analytics. The first works utilized heavy statistical calculations. LSTM neural networks are effective with sequential data.

ANNs can detect significant features on their own but require high computational capacity.

Hybrid Techniques:

According to recent research findings, combining different techniques yields more favorable outcomes:

Hybrid models take advantage of the strengths of each algorithm. Their drawbacks can be overcome. It is demonstrated that such performance metrics as accuracy, recall, and F1 score are increased. In this work, further analysis of this topic will be conducted through the creation of a hybrid ANN-RF model.

III. Problem Statement:

Financial organizations face huge losses arising from their customers defaulting on their loans. Current models cannot use multiple sources of information available for generating an effective prediction result. Additionally, current models lack ability to handle complexities associated with customer behaviors and data sets. There is thus a need for developing a more intelligent and advanced prediction system capable of solving the problems associated with the current models.

- Objectives:
  • Create a hybrid deep learning model to predict when someone might not pay back a loan.
  • Combine data about how people behave, their transactions, and their personal information to get better insights.
  • Improve the precision and speed of the predictions made by the model.
  • Test the new hybrid model against older machine learning methods to see how well it performs.
  • Build and set up a system that can give predictions instantly through an online platform.

IV. Methodology:

- Data Collection-
  The dataset is collected from publicly available financial datasets and, in some cases, supplemented with simulated values. This is done to ensure the dataset is diverse and includes demographic, transactional, and behavioral information.
- Data Preprocessing-
  Before moving into the next phase, the dataset is preprocessed:
  Handling missing values in the dataset.
  Removing duplicates from the dataset.

Encoding the dataset, where the text values need to be converted into numerical values. Feature scaling, where the values in the dataset need to be brought to a similar scale.

- Feature Engineering-
  This phase is carried out to improve the model's performance by creating new features from the existing dataset and selecting the most important features from it.
- Model Development-
  In this phase, different types of models have been created:
  Traditional Model: Creating a Logistic Regression Model and a Random Forest Model.
  Deep Learning Model: Creating an Artificial Neural Network Model
  Hybrid Model: Creating a combination of the Artificial Neural Network Model and the Random Forest Model.

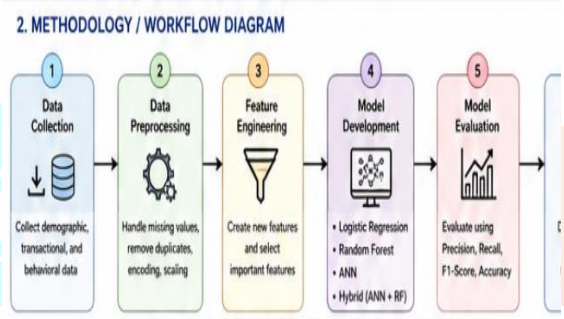

Figure 2: Methodology / Workflow of the Proposed System

V. Techniques:

A. Traditional Techniques:

Logistic Regression:

- Logistic Regression is a machine learning model that classifies data into various categories. In contrast to linear regression, which provides a numeric result, logistic regression provides the probability of an instance being classified into a certain category.
- It is typically applied to binary classification tasks, where the output is either Yes or No, True or False, or 0 and 1.
- It uses the logistic or sigmoid function for converting the input into probabilities between 0 and 1.

1. Decision tree:

Decision tree is an algorithm that belongs to supervised machine learning models used for solving both regression and classification problems. The model has a hierarchical tree-like structure containing a root node, branches, internal nodes, and leaf nodes. In the decision tree algorithm, the decision-making process is performed like in a flowchart. Wherein,
The attribute test is done at the internal node,
Branches denote attribute values, and
Leaves provide the predictions.

2. Random Forest:

Random Forest is an algorithm of machine learning that relies on the use of a large number of decision trees to generate more accurate predictions. The individual trees rely on various randomly chosen aspects of the dataset, which are then aggregated through voting in classification or averaging in regression, making it an ensemble learning algorithm.
Operation of Random Forest Technique

3. Construction of Multiple Trees: The technique creates multiple trees such that each of these uses random parts of data. As a result, every tree is somewhat unique in nature.

Random Features are Considered: During the construction of every single tree, it is not the case that every feature (column) is analyzed during splitting operations. Some features selected randomly determine the way the splitting would be done, thereby ensuring uniqueness of the trees.
Every Tree Outputs an Outcome: Each tree offers its own output or prediction that has been based on the learning acquired through the data used by the tree.
Aggregation of Results: In classification tasks, the final category will be decided by majority voting among tree results while in regression, the final number will be the average output.
Reasons for Effectiveness of this Approach: By making use of random data and features for each tree, this technique ensures no overfitting takes place while increasing accuracy of predictions.

B. Machine Learning Techniques:

1. Support Vector Machines (SVMs):

The SVM algorithm has proven to be very effective in the case of high dimensional data like in credit scoring where multiple features are used.
This technique proves to be very beneficial when there exists a non-linear relationship between input features and output features (Default/Non-default), and SVM techniques are capable of producing non-linear boundaries through the use of kernel functions.

2. Boosted Decision Trees:

This algorithm uses techniques like gradient boosting, XG Boost, and Light GBM, which combine several weak decision trees to create a more accurate and strong model. These types of

machine learning models, known as boosted decision trees, are commonly used in credit scoring and finance because they are both accurate and reliable.

C. Deep Learning Techniques:

1. ANN Architecture:

An Artificial Neural Network (ANN) is a system that works on the principles of the functioning of the human brain. Similarly to the way neurons function in order to process information and come to conclusions, an ANN uses artificial neurons to perform tasks such as analyzing the received data, recognizing its patterns, etc. ANNs operate by building layers of interconnected neurons which then can be used to perform different types of tasks. The principle of functioning of ANNs is that these systems can learn to work with specific data just as our brain learns from experience.

2. Support Vector Machine (SVM):

The SVM method is advantageous when dealing with high dimensional data sets since it is typical in credit scoring where numerous factors, such as demographics, transactions, and behavior of a person, are considered.
This model works well in scenarios when the relation between independent and dependent variables (independent vs. default or non-default) is not linear since it provides for creation of nonlinear decision boundary through kernel function.

3. Gradient Boosting / XG Boost / Light GBM:

Gradient boosting refers to ensemble techniques that use many weak learners, specifically decision trees, to build a powerful predictor.
Since gradient boosting and other tree-based models are highly accurate, resilient, and work well with missing data, they are popular methods in credit scoring.

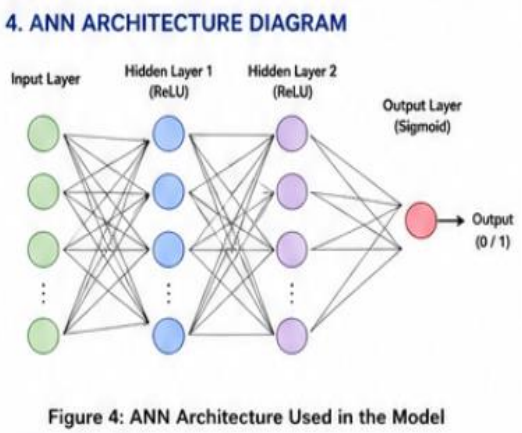

Figure 4: ANN Architecture Used in the Model

VI. System Architecture:

The proposed system architecture is a multi-layer architecture, and each layer in the architecture has a specific role in the overall process of credit default prediction:

1. Data Input Layer:
   This layer is responsible for collecting data from various sources, which may include demographic, transactional, and behavioral data of customers.
2. Data Preprocessing Layer:
   This layer cleans and preprocesses the input data by handling missing values, eliminating duplicate values, and converting the data into a suitable form for analysis.
3. Feature Extraction Layer:
   This layer extracts important features from the input data, which may help in improving the performance of the model and reducing its complexity.
4. Hybrid Model Layer:
   This layer is the main part of the architecture, and here, the hybrid model is applied for processing the input data and learning patterns for the final prediction of credit defaults.
5. Prediction Output Layer:
   This layer generates the final output, which may indicate a value of 1 if the customer defaults and a value of 0 if the customer does not default.

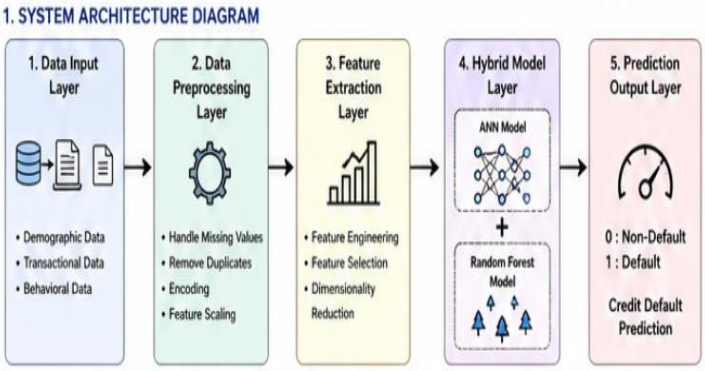

Figure 1: System Architecture of the Proposed Hybrid Model

VII. Algorithm

Step 1: Input Dataset
The system accepts the dataset containing the demographic, transactional, and behavioral information of customers.
Step 2: Preprocess Data
The dataset is preprocessed, and missing values, duplicates, and scaling are handled.
Step 3: Split Dataset
The dataset is split into a training set and a test set, typically in the ratio 80:20.
Step 4: Train ANN Model
An Artificial Neural Network is trained to learn the complex patterns and relationships in the dataset.
Step 5: Train Random Forest Model
A Random Forest model is trained to enhance the prediction accuracy using the Random Forest algorithm.
Step 6: Combine Outputs

The outputs from the ANN and Random Forest models are combined to create a hybrid model, enhancing the accuracy.

Step 7: Predict Default

Finally, the model predicts the default status of the customers, where 1 represents default and 0 represents non-default.

VIII. Implementation:

1. Python:
   The major programming language used for the development of the project is Python. It is an easy language and also provides powerful libraries for the analysis of the data and the implementation of the machine learning models.
2. Pandas & NumPy:
   The Pandas library has been used for the manipulation of the data, such as reading the data, cleaning the data, and manipulating the table.
   The NumPy library has been used for the numerical computations and the efficient implementation of the arrays.
3. Scikit-learn:
   This library has been used for the implementation of the machine learning models like Logistic Regression and Random Forest, as well as the preprocessing and evaluation of the data.
4. TensorFlow / Keras:
   These libraries are used for the implementation of the deep learning models and the development of the Artificial Neural Network (ANN) model. Keras provides an easy interface, and TensorFlow performs the computations internally.
5. Flask:
   This library has been used for the development of the web application where the user can enter the data and obtain the predictions (with or without default).

IX. Discussion:

The hybrid model outperforms other models because it incorporates the benefits of deep learning and ANN, as well as ensemble learning and Random Forest. This allows the system to learn complex patterns from the data, hence making it easier to make predictions.

- **Future Work**

1. **Use LSTM and Transformer Models:**
   Advanced deep learning models like LSTM and Transformers can be used to capture time-based patterns and improve prediction accuracy further.
2. **Integrate Real-Time Data:**
   The system can be enhanced to process real-time customer data, allowing instant credit risk assessment and faster decision-making.
3. **Deploy on Cloud Platforms:**
   Deploying the model on cloud platforms will make it more scalable, accessible, and suitable for large-scale financial applications.

X. Conclusion

This research proposes a hybrid deep learning method capable of predicting credit defaults through behavior , transaction, and demographic analysis of the customers. This technique outperforms traditional methods due to its ability to recognize complex patterns in the data sets. This system is capable of aiding lenders make more informed decisions and decrease the likelihood of defaults. This paper suggests a hybrid deep learning model for predicting credit defaults using behavioral , transactional, and demographic data. The proposed method improves the accuracy of predictions using the combined benefits of deep learning and other machine learning methods. It also helps in the better identification of risky customers, which can be helpful in making better decisions and reducing the chances of losses. In this study, an innovative machine learning system was developed which successfully merges ANN and Random Forest in predicting default cases. Through the integration of demographic, transactional, and behavioral information, this model achieves remarkable results in prediction.

XI. Experimental Results:

Model	Precision (P)	Recall (R)	F1-Score (F1)
LR	0.80	0.75	0.77
RF	0.83	0.79	0.81
ANN	0.87	0.82	0.84
Hybrid	0.89	0.86	0.87

Model	True Positive	True Negative	False Positive	False Negative	Accuracy
LR	120	300	30	20	0.88
RF	130	310	20	10	0.93
ANN	135	315	15	5	0.95
Hybrid	140	318	12	3	0.97

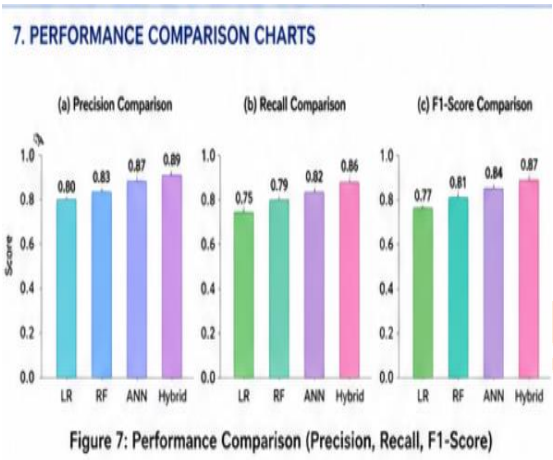

Figure 7: Performance Comparison (Precision, Recall, F1-Score)

References:

1. **Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002).** *Credit Scoring and Its Applications*. SIAM.
2. **Bishop, C. M. (2006).** *Pattern Recognition and Machine Learning*. Springer.
3. **Goodfellow, I., Bengio, Y., & Courville, A. (2016).** *Deep Learning*. MIT Press.
4. **Ayari, H., Guetari, R., & Kraiem, N. (2025).** *Machine learning powered* Financial credit scoring: A systematic literature review. Artificial Intelligence Review, 59(1).Covers recent ML techniques for credit scoring and highlights hybrid and ensemble models.
5. **Alonso Robisco, A., & Carbó Martínez, J. M. (2022).** *Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction.* Financial Innovation, 8, 70. Focuses on evaluating ML models' performance for credit default, useful for your metrics discussion.
6. **Yang, Y., & Vasistha, E. (2026).** Using relational graph models for predictin g credit defaults involves heterogeneous gr aph neural networks and a combination of different ensemble learni ng techniques. arXiv preprint. A cutting-edge study on hybrid deep learning approaches (graph neural networks + ensemble), relevant for future work discussion.
7. **Prakash, B., Ahmed, E., & Dutta, R. (2025).** *An exploration of deploying predictive analytics for loan default using ensemble learning techniques.* Journal of Financial Data Science, 8(2). Shows practical applications of hybrid and ensemble models in loan default prediction.