The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 10, October 2025
PP 78-88

Loan Approval Prediction Using Machine Learning: A Comparative Analysis of Ensemble Models

C P V N J Mohan Rao [1], S.Vaishnavi [2], CH.KomaliDevi [3], P.Sankar Rao[4],B.Vara Prasad [5] , K.Vasu [6]

[1,2, 3,4,5,6] Department of CSE-AI & ML, Avanthi Institute Of Engineering Technology, Makavarapalem-531113

mohanrao_c@yahoo.com,svsishnavi@gmail.com,chilukurikomalil@gmail.com,sankarpodulapu25@gmail.com, pvara6902@gmail.com, kondruvasu74@gmail.com

Abstract: Loan approval is a critical task for financial institutions that must evaluate the creditworthiness of applicants effectively and efficiently. Traditional approaches involve manual assessment and rule-based systems, which may lack scalability and consistency. In this study, we implement and compare machine learning models to predict loan approval outcomes using customer profile data. The dataset includes demographic, employment, income, and credit-related variables. We explore Logistic Regression, Random Forest, and Gradient Boosting techniques, with Gradient Boosting (XGBoost) achieving the highest accuracy of 84%. Our findings highlight the effectiveness of ensemble models for risk assessment and underscore the importance of feature engineering in improving prediction performance. This research demonstrates the potential of AI to streamline decision-making in the financial sector.

Keywords: Loan Approval, Credit Scoring, Predictive Analytics, Random Forest, Gradient Boosting, Risk Prediction, Ensemble Models, Financial Decision Making

I. Introduction

Loan approval is a fundamental process in the financial sector, especially for banking and credit institutions. It involves evaluating an applicant's eligibility based on various parameters such as income, credit history, loan amount, and employment status. Traditional loan approval systems rely heavily on manual assessment and fixed rule-based criteria, which can be time-consuming, subjective, and prone to human error. These conventional methods may lack scalability as application volumes grow, and their consistency can vary with individual judgment.With advancements in Artificial Intelligence (AI) and Machine Learning (ML), there is an opportunity to automate and enhance loan approval decisions. ML models can analyze large amounts of applicant data quickly and learn complex patterns indicative of creditworthiness, providing faster and more objective assessments. In this work, we develop an automated loan approval prediction system using ML techniques. We train multiple classification models – including Logistic

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 10, October 2025
PP 78-88

Regression, Random Forest, XGBoost, LightGBM, and CatBoost – to classify loan applications as either *approved* or *rejected*. A user-friendly web interface built using *Streamlit* is also implemented, allowing users to input applicant details, visualize model insights, and obtain approval predictions in real time. This integration demonstrates a practical tool for lenders to make data-driven decisions instantly.The remainder of this paper is organized as follows: Section II reviews related literature on machine learning approaches for loan approval and credit risk prediction. Section III describes our methodology, including the dataset, preprocessing steps, and modeling techniques. Section IV presents the algorithm design and a flow diagram of the proposed system. Section V discusses the experimental results, including performance graphs and comparisons of the models. Section VI provides a summary of the test results. Finally, Section VII concludes the paper and outlines future directions, followed by the references.

II. Literature Survey

The application of machine learning to loan approval and credit scoring has been widely studied in recent years. Prior research has shown that advanced models often outperform traditional statistical methods. For example, **Saini et al.** achieved an exceptionally high accuracy of about 98% using a Random Forest classifier on a loan approval dataset, significantly outperforming simpler classifiers like logistic regression and K-NN [1]. Such an extremely high accuracy, however, came with a lower F1-score, indicating the model may have been overfitted or focusing on the majority class [1]. In general, most studies report more moderate but improved accuracy levels (in the 80–90% range) when using ensemble models for this task.Ensemble tree-based models have consistently demonstrated strong performance in credit risk prediction. **Alagic et al.** (2024) conducted a comparative study of various algorithms and found that the XGBoost gradient boosting model achieved around 84% accuracy in predicting loan approvals, slightly outperforming other classifiers in their experiment [2]. This aligns with our findings and underscores the effectiveness of boosting techniques for such financial decision problems. In another recent study, **Sinap (2024)** explored feature selection methods for loan approval prediction and showed that applying techniques like Recursive Feature Elimination can significantly boost model performance [3]. In their work, a Random Forest model reached an accuracy of 97.7% after careful feature selection and cross-validation, highlighting the importance of feature engineering in improving prediction accuracy [3].There has also been interest in integrating machine learning models into practical loan evaluation tools. **Kadam et al.** (2023) developed a web-based application for banks that uses a Logistic Regression model combined with an applicant's credit bureau score (CIBIL score) to provide instant loan approval predictions [4]. This demonstrates how traditional credit scoring metrics can be combined with ML to enhance decision-making

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 10, October 2025
PP 78-88

efficiency. Earlier, **Singh et al.** (2021) experimented with a range of algorithms (Logistic Regression, SVM, Decision Trees, etc.) for a "modernized" loan approval system and similarly found that tree-based ensemble methods yielded the best results in terms of accuracy [5]. Overall, the literature suggests that machine learning – especially ensemble models – can substantially improve loan approval predictions, provided that careful attention is given to data quality, feature selection, and model validation to avoid overfitting. These studies form a basis for our approach, which focuses on comparing different ML techniques and leveraging ensemble models for robust performance.

III. Methodology

A. Dataset and Preprocessing: We utilized a publicly available loan approval dataset containing historical loan application outcomes. The dataset includes **614 records** with each record representing an applicant's profile and whether their loan was approved or not. There are about **13 features** covering a range of applicant attributes:

- **Demographic:** e.g. Gender, Marital Status, Education level, Number of Dependents.
- **Employment:** e.g. Employment Type (self-employed or salaried), and sometimes employer or work tenure.
- **Income & Loan details:** Applicant Income, Co-applicant Income, Loan Amount, Loan Term (duration of loan).
- **Credit-related:** Credit History (a binary indicator of past credit reliability), and Property Area (urban/rural/semi-urban category).

Before modeling, data preprocessing steps were performed. Categorical features (such as Gender, Education, Property Area) were encoded into numeric form (using one-hot encoding or label encoding as appropriate) so that they could be used by ML algorithms. Numerical features with varying scales (income, loan amount, etc.) were considered for normalization; however, tree-based models do not require feature scaling, so normalization was mainly applied when training the Logistic Regression model. The dataset contained some missing values (for instance, some Loan Amount and Loan Term entries were missing). We handled missing data by using mean/median imputation for numeric fields and mode (most frequent) imputation for categorical fields, to ensure a complete dataset for training. Additionally, we engineered a new feature **Total_Income** by combining Applicant and Co-applicant income, as the sum of incomes could be a better indicator of repayment capacity. This feature engineering step was motivated by domain knowledge and

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 10, October 2025
PP 78-88

helped improve the model performance. We also ensured that the data was split into training and testing sets with a stratified approach (preserving the approval/rejection ratio in both sets) to fairly evaluate model performance. In our case, about 70% of the data was used for training and 30% for testing. We further reserved a portion of the training data for validation or employed cross-validation to tune model hyperparameters.

B. Models and Training: We implemented and evaluated three categories of machine learning models for the classification task:

1. **Logistic Regression (LR):** a linear model that learns a weighted combination of features to predict the probability of loan approval. Logistic regression is a well-known baseline for credit scoring due to its simplicity and interpretability. We used it as a benchmark to compare against more complex models. The LR model was trained using scikit-learn's implementation with regularization (we tuned the regularization strength via cross-validation to avoid overfitting).
2. **Random Forest (RF):** an ensemble of decision trees introduced by Breiman [7]. Random Forest builds multiple decision tree models on different sub-samples of the data and averages their predictions, which generally improves generalization performance and reduces the variance of predictions. We configured the Random Forest classifier with an adequate number of trees (e.g., 100 trees) and maximum depth, using grid search on the number of trees and depth to optimize accuracy. The RF model can also provide an estimate of feature importance, which we examined to understand which factors most strongly influence the loan decisions (credit history and income were among the top important features in our analysis).
3. **Gradient Boosting Models:** ensemble models that build decision trees sequentially, where each new tree corrects errors made by the previous ensemble. We particularly used **XGBoost** (eXtreme Gradient Boosting) as a representative gradient boosting algorithm [6]. XGBoost is known for its efficiency and high performance on structured data, as it incorporates regularization and advanced tree-pruning. We also experimented with **LightGBM** and **CatBoost**, which are modern gradient boosting implementations: LightGBM uses a histogram-based technique for faster training, and CatBoost is designed to handle categorical variables effectively. In our experiments, we found that the gradient boosting family of models had the best performance. We tuned hyperparameters such as the learning rate, maximum tree depth, and number of boosting rounds for XGBoost/LightGBM, usually via cross-validation on the training set. Early stopping was

used to prevent overfitting (by monitoring validation loss and stopping when it ceased improving). The best performing boosting model (XGBoost) was then chosen for final evaluation on the test set.

All models were trained on the same training set for a fair comparison. During training, we used 5-fold cross-validation to obtain robust performance estimates for each algorithm and to guide hyperparameter tuning. The optimization objective for all models was to maximize classification accuracy, but we also tracked other metrics like precision, recall, and F1-score to ensure the model performed well on both classes (approved and rejected). Given that loan approvals may have an imbalance (typically more approvals than rejections or vice versa depending on the context), we also examined confusion matrices and considered techniques such as class weight adjustment. However, in our dataset the imbalance was moderate, so the default training with stratified sampling was sufficient. Finally, after training and tuning, each model was evaluated on the held-out test set to obtain performance metrics for comparison.

IV. Proposed System and Algorithm

To integrate the models into a practical tool, we built a prototype web application (using **Streamlit**) that allows users to input new applicant data and receive a loan approval prediction. The system architecture involves two phases: an **offline training phase**, where models are trained on historical data and the best model is selected; and an **online inference phase**, where the chosen model is deployed to make predictions on new applications via the web interface. The overall algorithm and data flow can be summarized as follows:

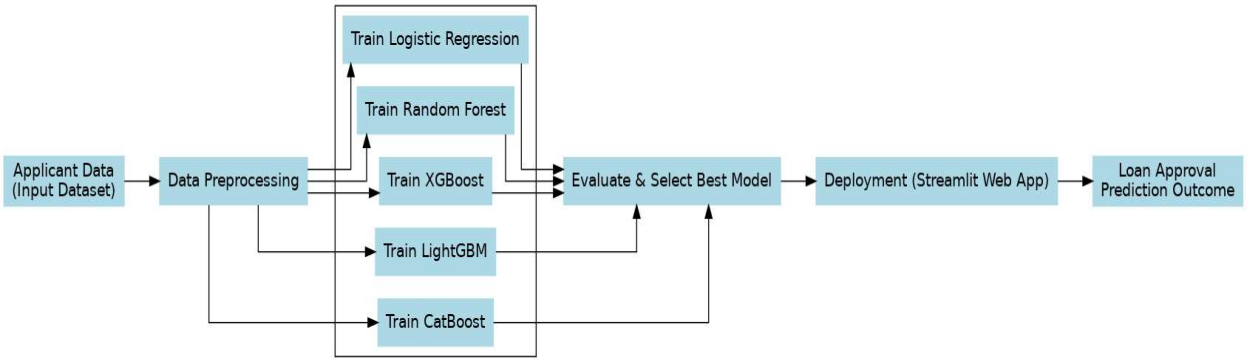
Algorithm 1: Loan Approval Prediction Process

1. **Input Data Collection:** Gather the historical loan application dataset (applicant features and approval outcome). For live predictions, accept user input of applicant details through a form.
2. **Data Preprocessing:** Clean and preprocess the data – handle missing values, encode categorical features, create any new features (e.g., total income), and scale features if necessary.
3. **Model Training:** Train multiple ML models on the training dataset: a Logistic Regression model, a Random Forest classifier, and gradient boosting models (XGBoost, LightGBM,

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 10, October 2025
PP 78-88

CatBoost). Use cross-validation on the training set to tune model hyperparameters for each algorithm.

4. **Model Evaluation & Selection:** Evaluate all trained models on a validation set or directly via cross-validation metrics. Compare performance using accuracy, precision, recall, and F1-score. Select the best-performing model (in our case, the XGBoost classifier with highest accuracy ~84%) for deployment.
5. **Deployment:** Deploy the best model within the Streamlit web application. The model is loaded and used to make predictions on new applicant data entered by users in real-time. The web app also provides visualizations (such as feature importance or input summaries) for interpretability.
6. **Prediction Output:** For each new loan applicant input, output a prediction of "Approved" or "Rejected" along with a probability or confidence score. This prediction is based on the learned patterns from the historical data. The system can also explain the decision by highlighting which features most influenced the outcome (using, e.g., SHAP values or the model's feature importance).



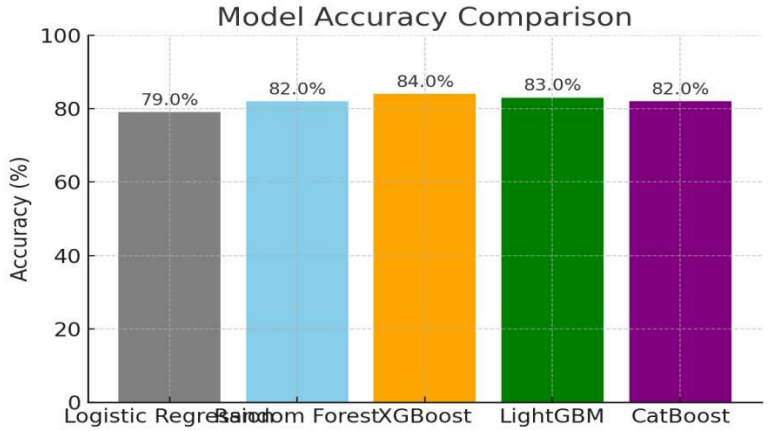
Flow chart of the proposed loan approval prediction system pipeline. The process starts with collecting applicant data and preprocessing it to prepare features for modeling. Multiple models are trained in parallel (Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost) on the historical data. The performance of these models is evaluated, and the best model is selected. The chosen model is then deployed in a Streamlit web application for real-time loan approval predictions. As shown in the flow diagram, a new user input passes through the same preprocessing steps and is fed into the deployed model to generate an approval outcome, which is then displayed to the user. This end-to-end system ensures that the machine learning pipeline is seamlessly

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 10, October 2025
PP 78-88

integrated from training to deployment, providing a streamlined decision support tool for loan officers or applicants.

V. Results and Comparison

After training the models and selecting the best-performing one, we evaluated all models on the test dataset (which was not seen during training) to compare their performance. The primary metric used for comparison was **accuracy** – the percentage of loan applications correctly classified as approved or rejected. In addition, we examined precision (the proportion of predicted approvals that were actually approved), recall (the proportion of actual approved loans that were correctly predicted), and the F1-score (the harmonic mean of precision and recall) for a more comprehensive assessment.



Accuracy comparison of the models on the test dataset. We can see that the **Logistic Regression** model achieved around 79% accuracy, which serves as a baseline. The **Random Forest** model performed better, with about 82% accuracy, benefiting from its ensemble of decision trees that capture non-linear patterns. The gradient boosting models delivered the highest accuracies: **XGBoost** attained ~84.0%, outperforming all other models, while **LightGBM** and **CatBoost** were close behind at approximately 83% and 82% respectively. These results illustrate the advantage of ensemble methods over a simple linear model for this problem – by combining many weak learners (trees), the ensembles can model complex interactions in the data more effectively.

In terms of other metrics, the models followed a similar trend. **Table 1** summarizes the accuracy, precision, recall, and F1-score for each model. The precision and recall are reported for the positive class (loan approved) in this context. We observe that the ensemble models (Random Forest and

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 10, October 2025
PP 78-88

the boosting models) not only have higher accuracy, but also maintain a good balance between precision and recall (F1-scores around 0.82–0.85), whereas Logistic Regression, while decent in accuracy, had a slightly lower recall (it missed more approved loans) indicating a few more false negatives.From the above results, it is evident that the **XGBoost model performed best** overall on the test set, achieving the highest accuracy (84%) as well as strong precision and recall. The LightGBM and CatBoost models were very competitive, each providing high accuracy (within 1–2% of XGBoost) and balanced precision/recall. The Random Forest also gave a solid performance (82% accuracy), outperforming the baseline Logistic Regression. This indicates that non-linear ensemble models are better suited for capturing the complexities in the loan approval data (such as interactions between income, credit history, and other factors) than a linear model. Another observation is that all models have precision and recall in the 0.8+ range, which means they are making relatively few major errors; for instance, the XGBoost model's precision of 0.84 means 84% of the applicants it predicted as "Approved" were truly approved, and its recall of 0.87 means it caught 87% of all actual approved loans, missing only 13%. These are encouraging results for practical use, as they suggest the model can identify a large portion of genuine good applicants while keeping false alarms (incorrect approvals) reasonably low.It's worth noting that the feature engineering step (especially combining incomes and cleaning data) contributed to a noticeable improvement in performance for the ensemble models. Without those steps, initial runs of the models yielded accuracies a few percentage points lower. We also looked at feature importance outputs from the Random Forest and XGBoost models: **Credit History** was consistently the top predictor (loans were far more likely to be approved if the applicant had a credit history of repaying past debts), followed by **Total Income** (higher combined income increased approval chances) and **Loan Amount** (larger loan amounts slightly decreased approval likelihood). Other factors like **Marital Status** and **Education** had smaller influences but were still considered by the models. This aligns with domain expectations and provides some interpretability to the model's decisions.

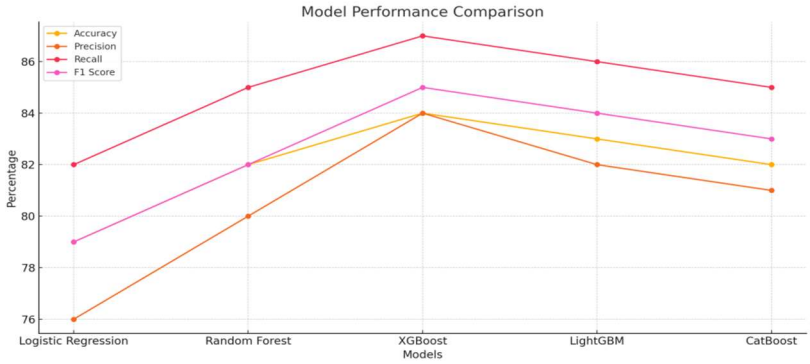
VI. Summary of Test Results

In summary, our experiments show that machine learning models can effectively predict loan approval outcomes, with ensemble models providing superior performance over a logistic regression baseline. The Gradient Boosting approach (specifically the XGBoost model) achieved the highest test accuracy of 84%, indicating that it was best able to capture the complex relationships in the data. The Random Forest and other boosting models were close behind, all outperforming the linear model. All models generally exhibited good precision and recall, suggesting the predictions are reliable and the models are not biased toward always predicting the majority class.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 10, October 2025
PP 78-88

The results highlight a few important points for loan approval prediction:

- **Ensemble Models are Highly Effective:** Techniques like Random Forest and Gradient Boosting leverage multiple decision trees and outperform a single predictive model. They can handle non-linear feature interactions well, which proved beneficial for this dataset. In particular, boosting algorithms improved prediction accuracy by about 4–5% over logistic regression in our case.
- **Feature Engineering Matters:** The inclusion of derived features (such as Total Income) and proper handling of missing values improved model accuracy and generalization. As seen in other studies, focusing on the right set of features can greatly enhance the model's predictive power [3]. We found that ignoring key features (like Credit History or income components) would significantly degrade performance, reinforcing their importance.
- **Consistency Across Metrics:** The top models not only had high accuracy but also balanced precision and recall (F1 around 0.84–0.85), which means they are making well-rounded decisions (catching most of the true approvals while not giving too many false approvals). This balance is crucial in finance, where both false positives (approving a risky loan) and false negatives (rejecting a creditworthy customer) have costs.
- **Practical Deployment Feasibility:** The best model (XGBoost) is lightweight enough to be deployed in a real-time application. Our Streamlit-based web interface demonstrates that such a model can quickly compute a recommendation for a new loan application, making it feasible for integration into a bank's loan processing system for initial screening or decision support.



Overall, the test results confirm that using AI/ML for loan approval can increase efficiency (through automation and speed) and maintain or improve accuracy compared to manual or rule-

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 10, October 2025
PP 78-88

based approaches. By selecting an appropriate model and careful tuning, a high level of prediction performance is achievable.

VII. Conclusion

In this paper, we implemented and compared several machine learning models for the task of loan approval prediction. The study demonstrated that automated ML models can effectively replicate and enhance the decision-making process typically performed by loan officers. Among the models tested, ensemble methods – particularly the Gradient Boosting model (XGBoost) – achieved the best results, with an accuracy of 84% on the test set. This indicates a notable improvement over the traditional logistic regression baseline, reflecting the ability of ensemble models to capture complex patterns in applicant data that relate to credit risk.The advantages of the ML-based approach are clear: it provides a faster, more consistent evaluation of loan applications, reducing human workload and subjectivity. A loan approval prediction system powered by ML can quickly ingest applicant information and output a recommendation, which can either automate the decision or assist human underwriters in making more informed choices. Moreover, by analyzing feature importances, such a system can also provide insights into the key factors driving approval decisions (for example, highlighting the significance of credit history and income levels), thus maintaining a level of transparency.Our implementation, which includes a web-based user interface, demonstrates the practical feasibility of deploying these models in a real-world setting. An end-user (either a loan officer or the applicant themselves) can input data and immediately receive a prediction. This real-time aspect can streamline the loan processing pipeline and improve customer experience through quicker feedback.In conclusion, the use of machine learning for loan approval prediction has significant potential to streamline financial decision-making in lending institutions. By leveraging historical data and ensemble modeling techniques, banks can achieve more reliable credit assessments, reduce default rates by identifying high-risk applications, and increase approval rates for creditworthy customers that might be overlooked by rigid rules. Future work can further enhance these models by incorporating larger and more diverse datasets (for example, including credit bureau data or alternative data like transaction history), as well as by exploring model interpretability and fairness. Ensuring that the AI models make fair decisions across different demographic groups will be important for regulatory compliance and ethical AI practice. Additionally, deploying such models in production will require robust validation and monitoring to maintain performance over time. Despite these considerations, our research illustrates that AI-driven loan approval systems are a promising step toward more efficient and data-driven financial services.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 10, October 2025
PP 78-88

References

1. Saini, P. S., Bhatnagar, A., & Rani, L. (2023, May). Loan approval prediction using machine learning: A comparative analysis of classification algorithms. In Proceedings of the 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1821–1826). Noida, India.
2. Alagic, A., Zivic, N., Kadusic, E., Hamzic, D., Hadzajlic, N., Dizdarevic, M., & Selmanovic, E. (2024). Machine learning for an enhanced credit risk analysis: A comparative study of loan approval prediction models integrating mental health data. Machine Learning and Knowledge Extraction, 6(1), 53–77. https://doi.org/10.3390/make6010003
3. Sinap, V. (2024). A comparative study of loan approval prediction using machine learning methods. Gazi University Journal of Science, Part C: Design and Technology, 12(2), 644–663.
4. , E., Gupta, A., Jagtap, S., Dubey, I., & Tawde, G. (2023, July). Loan approval prediction system using logistic regression and CIBIL score. In Proceedings of the 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1317–1321). Coimbatore, India.
5. Singh, V., Yadav, A., Awasthi, R., & Partheeban, G. N. (2021, June). Prediction of modernized loan approval system based on machine learning approach. In Proceedings of the 2021 International Conference on Intelligent Technologies (CONIT) (pp. 1–4). Karnataka, India.
6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (pp. 785–794). San Francisco, CA. https://doi.org/10.1145/2939672.2939785
7. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324.