Wireless Personal Communications (2022) 123:785–812
https://doi.org/10.1007/s11277-021-09158-9

Credit Scoring Models Using Ensemble Learning and Classification Approaches: A Comprehensive Survey

Diwakar Tripathi[1] · Alok Kumar Shukla[2] · B. Ramachandra Reddy[3] · Ghanshyam S. Bopche[4] · D. Chandramohan[5]

Accepted: 16 September 2021 / Published online: 1 October 2021

Abstract
Credit scoring models are developed to strengthen the decision-making process specifically for financial institutions to deal with risk associated with a credit candidate while applying for new credit product. Ensemble learning is a strong approach to get close to ideal classifier and it strengthens the classifiers with aggregation of various models to obtain better outcome than individual model. Various studies have shown that heterogeneous ensemble models have received superior classification performances as compare to existing machine learning models. Enhancement in the predictive performance will result great savings of revenues for financial institution. And, in order to provide the higher stability and accuracy, ensemble learning produces commendable results due to their inherent properties for improving the effectiveness of credit scoring model. So, this study presents a comprehensive comparative analysis of nine ensemble learning approaches such as Multiboost, Cross Validation Parameter, Random Subspace, Metacoast, etc. with five classification approaches such as Partial Decision Tree (PART), Radial Basis Function Neural Network (RBFN), Logistic Regression (LR), Naive Bayes Decision Tree (NBT) and Sequential Minimal Optimization (SMO) along with various ensemble classifiers frameworks arranged in single and multi layer with various aggregation approaches such as Majority Voting, Average Probability, Maximum Probability, Unanimous Voting and Weighted Voting. Further, this study presents the impact of various combinations of classification and ensemble approaches on six bench-marked credit scoring datasets.

Keywords Classification · Credit scoring · Ensemble classifiers · Ensemble learning

1 Introduction

Credit scoring is a widely adopted mathematical and statistical approach to evaluate the risk associated with an applicant applied for credit items. Basically, these mathematical and statistical approaches consider the applicant's credentials and applicants' historical data for estimation of the risk [1]. As indicated by Thomas et al. [2] "Credit scoring is a set

✉ Diwakar Tripathi
diwakarnitgoa@gmail.com

Extended author information available on the last page of the article

of decision models and their underlying techniques that aid credit lenders in the granting of credit", and its execution improves profitability of credit industries [3]. It efforts to separate the consequences of different candidates' attributes dependent on abnormal conduct and avoidances. Foremost emphasis of credit scoring is to preference whether a credit candidate can be measured as financially trustworthy "as credit product can be issued to applicant" or non-trustworthy "as credit product can not be issued to applicant" groups. Credit represents to the amount that can be issued to applicant by a financial organization and it is calculated by a model based on client's testimonial like salary, property and so on. Several rewards of credit scoring for credit industries integrate as "reducing credit risk", "making managerial decisions" and "cash flow improvement". Periodically, financial organizations succeed it in numerous stepladders such as "application scoring", "behavioural scoring", "collection scoring", etc. [4].

Application scoring provides an assistance for characterizing the new credit applicants for the assessment of the legitimacy or suspiciousness. That assessment is conducted in view of social, money related, and other data related to a credit applicant which are collected at the time of application. Behavioural scoring is similar as application scoring, but it provides assistance for dynamic portfolio administration progressions by inspecting the changes in behaviours on existing customers. Collection scoring classifies the present customers into several clusters based on deviation in their previous and current behaviours specially towards their purchasing behaviours. According to their group belongingness such as progressively, moderate, no, etc., banking system puts attentions on those groups [5].

Credit application scoring is an approach to arrange the credit candidate that it has a place with either authentic (financially trustworthy) or suspicious (financially non-trustworthy) group based on its accreditations. Enhancing the prescient execution of credit scoring model uncommonly candidates with non-trustworthy group will have incredible effect for financial institution [6]. Various researcher have considered it as binary class classification problem. From the literature, it is observed that individual classifiers show only moderately good performance as compared to ensemble classifiers [7–9]. Along with the ensemble framework, ensemble learning approach is also another way to improve the classification performance by selecting the appropriate training samples such as bagging, boosting etc. [10]. However, a classifier can not accomplish well with most of the datasets. Generally, a classifier can accomplish well with a specific dataset. Consequently, ensemble classifier is a robust and strong technique to get close to the optimum classifier for any dataset [11].

1.1 Motivation

As per the data released by Reserve Bank of India (RBI) [12] about credit card holders, there is approximately with 16% raise in number of card holders in year 2016 and total is 24.5 million. The card holders in the course of years 2012-2016 are as shown in Fig. 1. From the Fig. 1, it is clearly visible that every year there is a growth in number of credit card holders.

Together with credit cards, a verity of loans such as personal, vehicle, etc. are also obtainable by financial organizations. On account of abundant amount of new applicants and existing card holders, credit scoring is problematic to accomplish manually or it necessitates an enormous number of authorities with subjective knowledge on consumers behaviours. Nowadays, credit scoring is no longer restricted to manage an account or credit

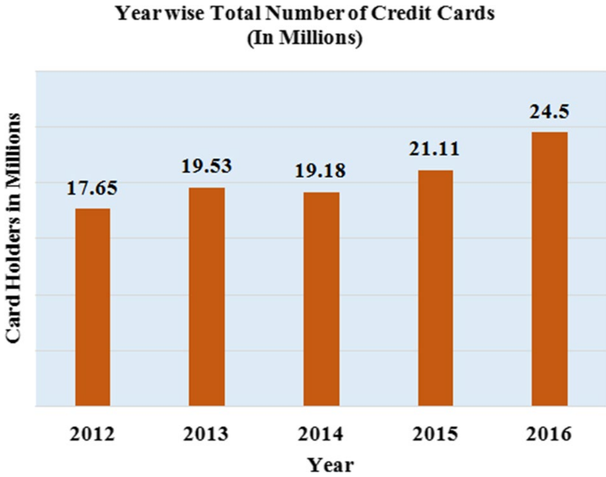

Fig. 1 Rise in number of credit card holders during the years (2012-2016)

businesses only, a variety of industries, including telecommunications, real estate, and so on are also utilizing the same to analyse clients' behaviour. So, limitations as mentioned earlier, machine learning is a way to solve the issue of manual credit scoring.

So, as per the machine learning perspectives, choosing the most suitable features and samples or elimination of redundant and irrelevant features and samples may enhance the predictive efficiency of credit scoring models. In literature, it is also displayed that ensemble classifiers framework also improves the classification performances and results are more robust as compared to results of individual classifier. But, it is not clear that which combination of ensemble learning with ensemble classifiers framework have the best way to apply for credit scoring or in other domain also. So, this study presents a comparative analysis with nine ensemble learning approaches such as Dagging, Metacoast, Multiboost, etc. with various classification approaches such as PART, RBFN, LR, NBT and SMO along with various ensemble classifiers framework with layered and single layer with various aggregation approaches such as Average Probability, Maximum Probability, and voting based approaches. And, its impact on six benchmark credit scoring datasets.

The rest of the study is organised in the following manner: Sect. 2 presents the summary of existing credit scoring models based on classification and ensemble classifications with its performances; Sect. 3 presents a brief introduction about classifiers which are utilized for ensemble framework; Sect. 4 presents the brief description of various ensemble learning, ensemble classifier framework and various aggregation approaches for aggregating the outputs predicted by base classifiers used for comparative results analysis; Sect. 5 presents descriptions about various credit scoring datasets utilized for comparative analysis, experimental results obtained in this study along with prior works and discussions; Sect. 6 presents the concluding remarks based on experimental outcomes.

2 Literature Survey

This section presents literature review on credit scoring specifically toward to machine learning perspective. By surveying the articles published in this domain, mainly the approaches are focused in two perspectives, first as credit score to evaluate the credit risk, and second as to transform the evaluation into binary-classification problems. A credit scoring model estimates a credit score against each credit product on the basis of customer's credentials, most of researchers have considered it as binary-classification problems and applied the machine learning approaches to find the hidden patterns from non-defaulter and defaulter customers credentials. Further, these patterns are utilized to identify new applicant as non-defaulter or defaulter. So, approaches intended for credit scoring are exhibited alongside their favourable circumstances and impediments, the significant demonstrating issues are examined from the machine learning perspective. Fig. 2 [3] portrays the recurrence of methodologies utilized by various researchers for credit risk assessment by considering more than 150 recent articles published in this domain. From the Fig. 2, it is visible that ensemble approach is the most well-known techniques utilized for credit scoring models. Some of the classifiers such as "Artificial Neural Network (ANN)", "Support Vector Machine (SVM)" and numerous additional classifiers have supported expressively to advance the credit risk prediction.

Several scholars have considered the SVM as classification approach for various applications because of its better regularization capabilities which results a reduced over-fitting problems and capable to handle non-linear data efficiently by considering an appropriate kernel function [13–18]. With some data pre-processing steps and SVM as classification of consumer loans applicants as default or non-default cases has been presented [19]. Numerous classification approaches namely: "Multi-layer Perceptron", "Mixture-of-experts", "Radial Basis Function", "Learning Vector Quantization", and "Fuzzy Adaptive Resonance" etc. are utilized for evaluating performance of classifiers and LR is observed as the most accurate approach for credit scoring [20]. As, ANNs have influential capability to learn and categorize complex non-linear associations between input and output [21]. In article [22], authors have integrated the external indicators for enhancing the predictive capability of ANN. As, the classification or predictions performance of ANNs depend on weights and biases associated to neurons at hidden layer, an evolutionary approach to get optimized weights and biases is presented in article [23]. As, consideration of appropriate features (or eliminating the redundant or irrelevant feature) can affect predictive capability of ANNs, an evolutionary approach to get optimized feature subset are presented in articles [24, 25].

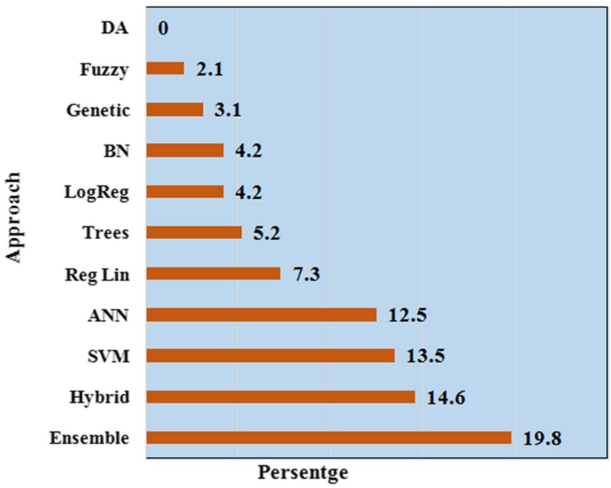

Fig. 2 Frequency of various methodologies utilized for credit scoring

Many classification methods for example SVM, ANNs and so on have been effectively employed for extracting the hidden knowledge from data of several problems. Though, a classifier can have better classification performance with a specific problem not with all problems. Ensemble learning is a strong approach to get close to ideal classifier and supports to the classifiers by adaptation of diverse models to acquire more accurate model as compared to individual model [11, 26]. Article [27], in this study six classification approaches namely: Naive Bayes (NB), LR, ANN, Discriminate Analysis (DA), *K*-Nearest Neighbours (KNN) and Classification Tree (CR) are applied. As the experimental figures are presented specifically classification performances, it was observed that ANN accomplishes better classification performance than the rest of five methods. Abellán and Castellano [10] have applied numerous ensemble approaches namely "Bagging", "Boosting", "Random Subspace", "Decorate" and "Rotation Forest" and revealed a comparative study for credit scoring. Similarly, Wang et al. [28] have also applied numerous ensemble learning approaches with LR, ANN, SVM and Decision Tree (DT). Amalgamation of base learner with Bagging achieves better than same with Boosting. Moreover, Amalgamation of base learner as "Stacking" and "Bagging" with DT achieved best outcomes towards to classification accuracy. From the experimental outcomes of as are presented in articles [10, 28] , it was exposed that normal models depending on individual classifiers or a basic mix of these classifiers in general displayed moderate execution. From the experimental results as are depicted in the article "A comparative study on ensemble classifiers for bankruptcy prediction" offered by Nanni and Lumini [29], it can be concluded that "Random Subspace" with "Levenberg Marquardt Neural Nets (LMNC)" have the best accomplishment when contrasted with rest of the methodologies. Zhang et al. [30] have applied "Vertical Bagging Decision Tree Model (VBDTM)" for credit scoring.

Towards to previous surveys in credit scoring models, Lin et al. [31] have presented a comprehensive survey on financial crisis prediction models specifically in Machine Learning perspective in terms of categorization of approaches as classification, ensemble and hybrid along with datasets utilized. Further, they have presented the classification accuracy achieved by these approaches on various credit scoring datasets. On various datasets, performance of the hybrid and ensemble classifiers provide more reliable conclusions. Lahasasna et al. [32] have offered a review on methods utilized for emergent credit scoring models with issues are deliberated particularly towards to machine learning point of view. Abdou and Pointon [33] have presented an article entitled "credit scoring, statistical techniques and evaluation criteria: a review of the literature". In this article, numerous performance measures with numerous statistical methods as are utilized by financial and banking professionals. Additionally, an evaluation between various statistical methodologies exhibited that complex procedures such as ANN and genetic programming, perform better than more conventional methodologies such as DA and LR, in terms of predictive performance.

Tripathi et al. [53] have offered a comparative study on numerous filter methods for feature selection and its influence on numerous classification and ensemble methods. As results are portrayed, STEP based feature selection with weighted voting based layered ensemble classifier has the best classification performance. Results of various ensemble frameworks in layered and non-layered manner are displayed in article [9, 49]. From the results, it is observed that layered approach with WV has improved performances as compared to its base classifiers employed for construction the ensemble framework. Authors in article [54] presented an approach for discrimination in between worthy and non-worthy debt customers based on the current refined feature selection methods to identify the most favourable features with relevant information. In addition, author deliberated numerous issues associated to applicability of feature selection approaches. Furthermore, deliberated about the problems that used to be insufficiently underlined in standard feature selection works. Multiple Population Genetic Algorithm based A hybrid approach (as HMPGA) is presented in article [55]. In this article, wrapper approach in association filter approaches to acquire significant prior information for initial populations setting of MPGA with characteristics of global optimization and quick convergence is presented to find optimal feature subset. Tripathi et al. [56] have offered an experimental result analysis on nine filter methods for feature selection and eight heterogeneous classification approaches and concluded that Unsupervised Discriminative Feature Selection (UDFS) has the best outcomes with most of the classification approaches.

Furthermore, we have gone through the published articles in this domain, and categorized those approaches into three categories as classification, ensembles and hybrid and found that Australian and German datasets are the mostly utilized datasets for experimental analysis. Experimental results specifically classification accuracy of respective approaches in respective categories (as approaches are categorized in three categories) along with dataset as "Australian Dataset (AUS)", "Japanese Dataset (JPD)", "German Categorical Dataset (GCD)", "German Numerical Dataset (GND)" and respective references are tabularized in Table 1. From the experimental outcomes as in Table 1, it is observed that Neighbourhood Rough Set (NRS) with Layered Weighted Voting (LWV) (with "Multilayer Feed Forward Neural Network (MLFN)", "NB" and "Quadratic Discriminant Analysis (QDA)" at layer first and in last layer "Distributed Time Delay Neural Network (DTNN)" and "Time Delay Neural Network (TDNN)") have the finest classification accurateness with Australian and German datasets respectively.

3 Classifiers

This section presents brief explanation about various classification methods specifically: PART, RBFN, LR, NBT and SMO applied in this survey for credit scoring data classification.

3.1 Partial Decision Tree

PART [57, 58] is an efficient rule based classification approach and it associates two approaches namely: C4.5 and ripper to evade their individual issues. In contrast to previous approaches, it doesn't consider global optimization for constructing the rule set. For creating a rule, it makes the use of pruned DT with present instances with leaf with prime exposure. Further, tree is discarded. The possibility of repeatedly construction of decision trees just to dispose of the majority of them, which are not as odd as it initially appears. Using a pruned tree to secure a standard instead of pruning a standard consistently by incorporating conjunctions with every one to avoids a tendency to over prune, which is a trademark issue of the "separate-and-conquer rule learner". By utilizing the "separate-and-conquer methodology" for elimination of the covered instances in conjunction with decision trees adds flexibility and speed. The key thought is to construct a partial decision tree instead of a fully explored one. To produce such a tree, the development and pruning tasks are incorporated so as to locate a "stable" sub tree that can be disentangled no further. When this sub tree has been discovered, tree building stops and a solitary standard is scrutinized off.

Table 1 State-of-the-art-approaches with predictive accuracy for credit scoring

Method	Dataset				References
	AUS	JPD	GCD	GND	
Classification approaches					
ANN	84.10	–	72.80	–	[34]
KNN	83.60	–	66.90	–	[34]
SVM-L	87.40	–	74.80	–	[34]
SVM-R	86.10	–	75.90	–	[34]
CART	85.90	–	55.90	–	[34]
J48	84.50	–	64.10	–	[34]
LR-R	86.20	–	75.40	–	[34]
RBFN	87.14	–	74.60	–	[20]
MLP	85.84	–	73.28	–	[20]
LVQ	82.97	–	68.37	–	[20]
Ensemble approaches					
RS+DT	88.17	–	78.52	–	[6]
CHE	88.10	88.70	79.00	–	[26]
CSA	88.98	87.88	77.72	–	[35]
RF+CTD	86.10	86.40	75.20	–	[10]
VBDTM	91.97	–	81.64	–	[30]
RS+LMNC	87.05	87.34	73.93	–	[29]
NNs	87.25	85.91	76.60	–	[36]
Bstacking	88.28	–	78.66	–	[37]
MSEC	87.40	87.00	78.30	–	[38]
Hybrid approaches					
SR+ANN	84.09	–	–	–	[39]
FS+SVM	86.76	–	76.84	–	[40]
GA+SVM	90.19	–	84.24	–	[41]
NRS+SVM	–	85.48	74.50	–	[42]
IFS	90.90	–	–	80.20	[43]
HGA-NN	–	–	78.90	–	[44]
SVM+GS	85.51	–	76.00	–	[45]
SVM+GS+FS	84.20	–	77.50	–	[45]
SVM + GA	86.90	–	77.92	–	[45]
NRS+SVM+GS	87.52	–	76.60	–	[46]
GA+NB	85.56	–	–	74.03	[47]
LDA+MLP	86.00	–	–	73.44	[47]
RS+TS+LR	–	86.40	–	–	[48]
LWV	92.58	89.16	85.82	85.62	[49]
NRS+LWV	95.39	–	86.47	–	[9]
FS+WV	87.32	87.98	77.12	–	[50]
MHNGA	87.54	87.20	76.82	–	[51]
GFSS	88.10	89.40	87.60	–	[52]

Grid search (GS), Consensus hybrid ensemble (CHE), F-score (FS), Consensus system approach (CSA)

3.2 Radial Basis Function Neural Network

RBFN is a three layers (as Input, Pattern and Summation) feed forward architecture, it forecasts the probabilities of input sample to the classes, and it estimates the probability by a linear combination of radial basis functions of the inputs and neuron parameters [59, 60]. Each adjacent layer (such as input-pattern, pattern-summation) are fully connected layer, with the number of neurons as the number of features, samples and classes in training dataset respectively. Each neurons in pattern layer is described by the redial basis function as in Eq. 1. Complete mathematical process of RBFN for prediction is as follows in Eqs. 1–4 and 1 presents the most common radial basis function.

$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma_i}} * exp^{-\frac{(x-\mu_i)' *(x-\mu_i)}{2\sigma_i^2}}, i = 1, 2, \dots, K \tag{1}$$

$$R = \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \dots & \phi_k(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \dots & \phi_k(x_2) \\ & & . & \\ & & . & \\ & & . & \\ \phi_1(x_M) & \phi_2(x_M) & \dots & \phi_k(x_M) \end{bmatrix}, T = \begin{bmatrix} T(x_1) \\ T(x_2 \\ . \\ . \\ . \\ T(x_M) \end{bmatrix} \tag{2}$$

$$W = \alpha^{-1} * R' * T, \qquad \alpha = [R' * R] \tag{3}$$

$$Y_i = W_i * \phi(x), \quad i = 1, 2, \dots, M \tag{4}$$

Various indicators used in Eqs. 1–4 are as follows, input vector as x, output of i^{th} hidden neuron as ϕ_i, center vector μ_i. R and R' indicate radial basis matrix and transpose of the matrix R which is generated as described in Eq. 2, $T(x_{(1...M)})$ describes the target value to corresponding training pattern. T is target vector of training dataset and α is a variance matrix. Where, Y_i is the i^{th} output which is weight sum of hidden neurons.

3.3 Logistic Regression

LR [61] is a predictive investigation procedure based on the concept of probability and it can be considered as an extraordinary case of linear regression models "with binary class classification, it violates normality assumptions of general regression models. LR indicates that a proper function of the fitted likelihood of the event is a linear function of the observed values of the available explanatory variables". The noteworthy favoured viewpoint of this philosophy is that it can make a clear probabilistic formulation of interpretation. Discriminant function analysis is basically the same as LR, and both can be utilized to respond to a similar research queries [62]. LR doesn't have the same number of presumptions and limitations as DA. Though, when DA assumptions are encountered, it is more dominant than LR [63]. Rather than LR, DA can be applied with minor sample size and homogeneity of co-variance, DA is gradually.

With k classes, n instances and m features, the parameter matrix B to be calculated and will be an *m*(k-1)* matrix by considering "squeeze" for optimization procedure. The probability for class j with the exception of the last class is as follows:

$$P_j(X_i) = \frac{exp(X_i * B_j)}{\sum_{j=1}^{k-1} exp(X_i * B_j) + 1} \tag{5}$$

The last class has probability

$$1 - \sum_{j=1}^{k-1} P_j(X_i) = \frac{1}{\sum_{j=1}^{k-1} exp(X_i * B_j) + 1} \tag{6}$$

The (negative) multinomial log-likelihood is thus:

$$\begin{aligned} L = & - \sum_{i=1}^{n} \sum_{j=1}^{k-1} Y_{ij} * ln(P_j(X_i)) \\ & + (1 - \sum_{j=1}^{k-1} Y_{ij} * ln(1 - \sum_{j=1}^{k-1} P_j(X_i)) + ridge * B^2 \end{aligned} \tag{7}$$

In order to calculate the matrix B for which L is minimised, "Quasi-Newton Method" is utilized to obtain the optimized values of the *m*(k-1)* variables.

3.4 Naive Bayes Decision Tree

DT [64] is a predictive modelling method. For learning, it constructs a decision tree from class-labelled training samples. In this model, "observations" and "corresponding target values" are characterized in the branches as "conjunctions of features" and "leaf nodes" respectively. In case of NBT, the leaf node is categorized by Naive Bayes with standard entropy as criteria for categorizing the continuous attributes to categorical, instead of considering the single class [65].

3.5 Sequential Minimal Optimization

SVM as classification approach has better performance but it has complex training and it requires expensive "third-party Quadratic Programming (QP) solvers" [66]. SMO as classification approach has capability to resolve QP of SVM with consideration of disintegrating the general issue into different sub-issues and by employing the slightest conceivable optimization approach at each progression with two Lagrange multipliers to locate the optimal qualities [67]. SMO is expressed in the dual form as follows in Eq. 8.

Let, $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, \dots, y_n\}$, where, $X$and Y represent training samples and target vector with n samples and x_i symbolizes as input vector and y_i symbolizes the class label of x_i.

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j, \tag{8}$$

Subject to: $0 \le \alpha_i \le C, \quad$ for $i = 1, 2, \dots, n,$

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \tag{9}$$

Where, C presents SVM hyper parameter and $K(x_i, x_j)$ is the kernel function, both supplied by the user and the variables are Lagrange multipliers.

4 Ensemble Approaches

Ensemble approaches are well recognized technique for procurement of highly accurate classifiers by mingling less accurate ones. Numerous techniques have been anticipated for developing the ensembles, and as per the learning mechanism, these techniques can be classified into two categories as ensemble learning and ensemble classifiers framework. Ensemble Learning refers to "The procedures employed to train multiple learning machines and combine their outputs, treating them as a committee of decision makers" with motivation that the committee (of base learners) decision may be more precise and robust than individual base learner [68]. Ensemble classifiers framework (Multi-Classifier Systems) focus on "The combination of classifiers form heterogeneous or homogeneous modelling backgrounds to give the final decision" [69–71]. Whereas, with ensemble framework same training set is supplied to train the various classifies and further the output of various classifies are aggregated by aggregation approaches for predicting the final output.

4.1 Ensemble Learning

In case of ensemble learning approach different sets from a training samples set are supplied to train the classifiers and output against a sample is aggregated by majority voting for predicting the final output. For generation of various training set, there are various approaches such as consideration of a subset, manipulation the training set, manipulation of input features and injecting randomness in training set [68]. Various ensemble learning approaches are explained as follows in Sects. 4.1.1-4.1.9

4.1.1 Bagging

Bagging acronym derived from "**B**ootstrap **AGG**regat**ING**" is an ensemble learning technique [72]. Let the training set D with n samples, further it generates m new training sets D_i "with *n'= (1 - 1/e) (approximately 63.2%)* unique samples" are chosen from training set, rest "*(n-n')*" are duplicated samples [73] and this is known as a bootstrap sample. Further, m models are fitted by utilizing aforementioned m bootstrap samples and combined by averaging the output and voting in case of regression and classification respectively.

4.1.2 Cross Validation Parameter

Cross Validation Parameter is a wrapper approach, which is considered as a black box with tunable parameters [74]. For tuning the parameters, training dataset is segregated into internal training and test sets, with dissimilar settings of the parameters. The setting with the most noteworthy evaluated esteem is picked as the last parameter set on which to run the enlistment calculation. It has two crucial components to the wrapper approach as search and evaluation component. Where first component recommends parameter settings and second component evaluates parameters as chosen in previous by executing the induction algorithm several times and receiving an optimized parameter as per the objective function, usually accuracy.

4.1.3 Adaptive Boosting

Adaptive Boosting [75] also involves Bootstrapping. In contrast to Bagging, Boosting considers some samples are utilized more regularly than others. Moreover, the Boosting makes the use of "weak learning algorithm" which is indicated conventionally as "WeakLearn". It fits a progression of weak learners on divergent weighted training data. It begins by foreseeing one of a kind dataset and surrenders measure to weight to each recognition. In the occasion that conjecture is erroneous using the essential learner, by then it puts higher weight to recognition which have been foreseen mistakenly, and this process is continued until threshold is not crossed. The threshold can be considered as the number of iterations or model's accuracy (or error rate).

4.1.4 Decorator

"Diverse Ensemble Creation by Oppositional Relabelling of Artificial Training Examples (Decorator)" is a process for constructing ensembles that directly generates miscellaneous propositions by means of supplementary artificially-generated training samples [76]. Both, some proportion of actual and artificial training samples are considered for training the classifier. For producing the artificial training data, with numerical features " mean, standard deviation and Gaussian distribution" and with nominal features " probability of occurrence of each distinct value with Laplace smoothing" are applied and labels for these produced training samples are chosen so as to differ maximally from the current ensemble's predictions. In each iteration, the number of artificial data samples to be subject to size of actual training samples.

4.1.5 Random Subspace

Random Subspace is an ensemble learning approach and it considers random samples with replacement of feature set similar to Bagging with motivation "individual learners should not over-focus on features that appear highly predictive or descriptive in the training set, but fail to be as predictive for points outside that set" [77]. So, it considers different set of features to train the different models and further, it aggregates the output of different models by majority voting.

4.1.6 Rotation Forest

To produce the training samples for a base classifier, the feature set is arbitrarily divided into K subsets, further it applies "Principal Component Analysis (PCA)" with training set [78]. All principal components are involved in order to preserve the inconsistent information in the data. In this manner, K subsets rotations take place to shape the new features for a base classifier. The main reassurance behind the rotation approach is to give confidence concurrently individual accuracy and diversity inside the ensemble. Diversity is advanced through the feature extraction for each base classifier.

4.1.7 Dagging

Similar to Bagging, Dagging also considers association of numerous trained models to derive a solitary model with contrast that Bagging utilizes bootstrapping approach and Dagging utilizes the disjoint set of samples to train the models [79].

4.1.8 Metacost

"A General Method for Making Classifiers Cost-Sensitive: MetaCost", is based on wrapping a "meta learning" stage around the error-based classifier in such a way that the classifier effectively minimizes cost while seeking to minimize zero-one loss [80]. The conditional risk $R(i|z)$ as in Eq. 10 is the expected cost of predicting that z belongs to class i by utilizing Bayes optimal prediction for z. And, for relabelling the training samples with "optimal" classes, it estimates by class probabilities $P(j|z)$.

$$R(i|x) = \sum_{j} P(j|x)C(i,j) \tag{10}$$

4.1.9 MultiBoost

MultiBoosting is an extension to the extremely powerful AdaBoost strategy for edging decision committees. It is a hybrid approach by fusing AdaBoost with Wagging [81] with motivations as follows "1. Bagging mainly reduces variance, while AdaBoost reduces both bias and variance and there is evidence that Bagging is more effective than AdaBoost at reducing variance", "2. Wagging [82] is an alternative of Bagging, which requires a base learning calculation that can use preparing cases with differing weights. As the mechanisms differ, their combination may out-perform either in isolation".

4.2 Ensemble Framework

There are numerous classification algorithms, however there is no particular method to foresee which classifier will deliver the best outcomes on a particular dataset. An ensemble of classifiers has the capacity to deliver close optimal outcomes on each dataset. An ensemble framework can be amassed in two ways homogeneously "association of same type of base classifiers" or heterogeneously "association of diverse type of base classifiers". Further, either heterogeneous or homogeneous base classifiers can be aggregated in single-layer or multi-layer. Ensemble frameworks with single layer and multilayer are framed as in Fig. 3 and Fig. 4 individually. A multi-layer ensemble classifier framework permits adaptation from multiple points, unlike a single layer classifiers [83], as the diverse classifiers at different layers can utilize diverse features set at each layer and the classification tasks can be more refined. The computational complexity of the multi-layer framework is reduced by isolating it into a multi-layer framework. The foremost purpose behind employing a multi-layer ensemble framework is that, when the classifier makes a decision, it isn't reliant on only a solitary classifier's choice, in any case, rather, requires all classifiers to take an interest in the basic leadership process by

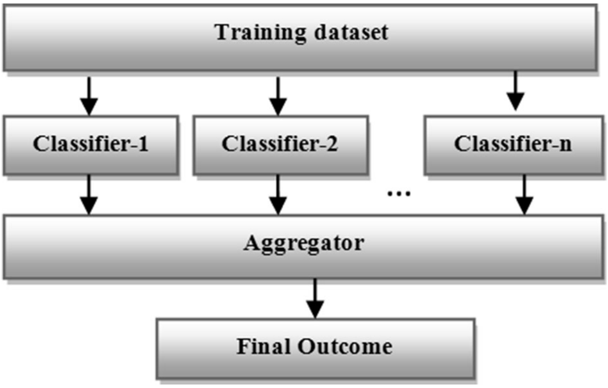

Fig. 3 Architecture for single layer ensemble layer framework

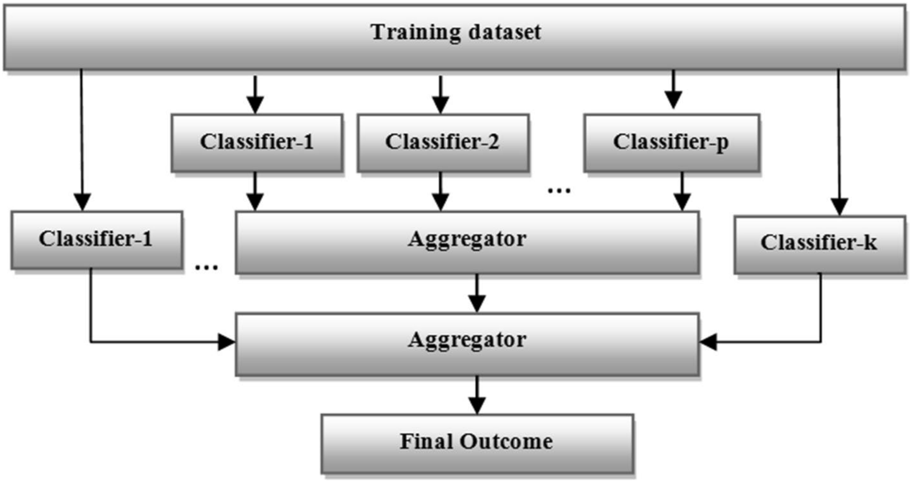

Fig. 4 Architecture for multi layer ensemble layer framework

conglomerating their individual expectations. Thus, this technique outflanks the base classifiers.

In this study, five heterogeneous classifiers such as PART, RBFN, LR, NBT and SMO are aggregated into single layer and multilayer ensemble classifier framework. Because from the literature, it was observed that ensemble classifier with homogeneous base classifiers has best classification performances towards to a particular class. In case of single layer ensemble framework as in Fig. 3, all output predicted by five classifiers are aggregated by aggregator, and it will be the final output against that sample. In case of multilayer ensemble framework as in Fig. 4, output predicted by first p classifiers are aggregated by aggregator, and output of aggregator and remaining output predicted by $k=n\text{-}p$ classifiers are forwarded to aggregator at second layer, output of aggregator at second layer will be the final output against that sample. Various aggregation approaches are explained as follows in Sects. 4.2.1-4.2.3.

4.2.1 Majority and Unanimous Voting

Majority voting and unanimous voting approaches as aggregator, it applies voting on outputs predicted by base classifiers and final output will be the class which has the highest votes [70] and the class by all base classifiers [84] respectively. In case of unanimous voting approach, it may have the best and most robust towards to a particular class but it will have the worst performances towards to other classes. This is somewhat obfuscating a very straightforward procedure, and it calculates the total votes obtained by for class j as summation by $\sum_{t=1}^{T}$. Further, it makes the most of the sum " presumably with a coin flip for tie breaks" as in Eq. 11 [68].

$$\sum_{t=1}^{T} d_{t,j} = \max_{j=1}^{C} \sum_{t=1}^{T} d_{t,j} \tag{11}$$

4.2.2 Max and Average Probability

Max and Average Probability as aggregator, both aggregates the outcomes attained by the base classifiers by consideration of approximating the summation of the "maximum of the posterior probabilities" and "average posterior probability" for each class over all the classifier outputs respectively. Max and Average Probability as aggregators can be calculated by the equations as follows in Eqs. 12 and 13, respectively [85].

$$\max_{j=1}^{C} P(w_j|x_i) = \max_{k=1}^{m} \max_{j=1}^{C} \sum_{t=1}^{T} P(w_k|x_i) \tag{12}$$

$$Med_{j=1}^{C} P(w_j|x_i) = Med_{k=1}^{m} \max_{j=1}^{C} \sum_{t=1}^{T} P(w_k|x_i) \tag{13}$$

4.2.3 Weighted Voting

Weighted voting as aggregator accepts the outputs by the base classifiers and are aggregated as weighted sum. For assigning the weight to base classifiers, is reversely proportional to misclassification rate [70]. Each aggregator aggregates the output predicted by the associated classifiers using Eq. 14. For assigning the weights to base classifiers, initially equal weights are assigned to each base classifier. Further, it is updated as in Eq. 15 [70]. This procedure will be continued upto n iterations, and at last the mean will be considered as final weights to the respective classifiers.

$$O = \sum_{i=1}^{P} W_i * X_i \tag{14}$$

$$W_{ij} = \frac{(1 - Er_{ij})}{\sum_{j=1}^{P}(1 - Er_{ij})} \tag{15}$$

With P base classifiers and W_i and X_i as weight and predicted output of the i^{th} classifier [86]. Where, W_{ij} and Er_{ij} symbolize the weight and classifier's error of j^{th} classifier in i^{th} iteration respectively.

5 Results and Discussion

According to the objective of this study, this section is partitioned into three sub-sections: first sub-section introduces the dataset and performance measures, second as result analysis of various classification and ensemble framework with various ensemble learning approaches, at last comparisons with other state-of-the-art techniques are conducted in the third sub-section.

5.1 Credit Scoring Datasets

To validate the effectiveness of credit scoring models, six most popular (as most of the published article have utilized these datasets to show the effectiveness of their models) benched-marked credit scoring datasets as Taiwan, Bank-marketing, German-categorical, German numerical Australian and Japanese datasets are chosen. Datasets specifically: Australian and German categorical are the furthermost widespread datasets and approximately 80% articles have utilised these two datasets for experimentation. Comprehensive explanation of aforementioned datasets attained from UCI data repository [87] used in this article are tabularized in Table 2. All the datasets are real world credit scoring datasets and are related with different credit products application such as loan, credit card etc. and because of confidentiality, some of the feature values are transmuted by figurative representation.

Taiwan dataset is of an important bank in Taiwan. In this dataset targets are credit card holders of the bank [27], and features are completely numerical. First five features are about personal status of candidate, next 18 features are about last 6 months payment status (as paid on time, delay or partial payment), amount of billing statement and amount of previous payment. Bank Direct Marketing (Bank-marketing) dataset [88] is of direct marketing campaigns conducted by a Portuguese financial institution of 45211 applicants with 16 different applicants' status related to personal, financial etc. and these details are collected over the phone call. Further, based on credential financial institution have categorized the applicants into two groups such as "yes" and "no" (as creditworthy and non-creditworthy

Table 2 Detailed description about benched-marked credit scoring datasets utilized for comprehensive comparative analysis

S. No	Dataset	Number of samples	Ratio of class-1/class-2	Number of features	Ratio of features categorical/ Numerical
1	Taiwan	30000	23365/6635	23	0/23
2	Bank-marketing	4521	4000/521	16	9/7
3	German-categorical	1000	700/300	20	13/7
4	German-numerical	1000	700/300	24	0/24
5	Japanese	690	307/383	15	9/6
6	Australian	690	307/383	14	8/6

group). As it is quite big dataset, so same institution has provided a slighter dataset to assessment more computationally challenging machine learning approaches (e.g., SVM) with 4521 samples with 16 features as bank dataset. Bank dataset is produced by the same institution by considering 10% samples of each class. German-categorical dataset [89] are of loan applicants in a bank in Germany and have 1000 samples with 20 features which defines the applicant's history with the ratio of 7:3 creditworthy and non-creditworthy applicants. For algorithms that need numerical attributes, Strathclyde University produced the file "German Data-numeric". This document has been altered and a few marker factors added to make it appropriate for calculations which can't cope with categorical variables. Both the German (categorical and numerical) credit scoring datasets are loan approval datasets. Australian [90] and Japanese [91] datasets are associated to credit card aspirants and both have categorical and numerical features.

In this study, we have considered accuracy for comparative result analysis. As in this article, we have considered credit scoring approaches as binary class classification problem (creditworthy and non-creditworthy cases). And, accuracy measures the percentage of creditworthy and non-creditworthy cases are classified correctly. Mathematically, it is stated as follows in Eq. 16.

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{16}$$

Where, T_P, T_N, F_P and F_N are the indicated as "True Positive", "True Negative", "False Positive" and "False Negative" respectively.

5.2 Result Analysis

This section presents the results obtained by various approaches in six credit scoring dataset in terms of classification accuracy. Main motive of this study is to present a comprehensive comparative result analysis on various ensemble learning approaches and ensemble frameworks and combination of both. As in this study, MV is utilized as aggregation approach which has the limitation that there must be odd number of classifier. And, in ensemble classifiers with homogeneous base classifiers have better classification performance towards to a specific class. So, in this study, we have considered five heterogeneous classifiers namely: NBT, PART, MLP, LR and SMO classifiers as base classifiers. And, various ensemble learning approaches such as Bagging (Bagg), CVparameter (Cvpar), Adaboost (Adab), Decorator (Deco), Subspace (Subs), Rotation Forest (ROFo), Dagging (Dagg), Metacoast (Metac), Multiboost (Multib) with all aforementioned classifiers are considered. Along with ensemble learning, ensemble framework in single and multilayer with various aggregation approaches with aforementioned five heterogeneous classifiers namely: Majority Voting (MV), Average Probability (AvgPro), Maximum Probability (MaxPro) Unanimous Voting (UV) and Weighted Voting (WV) are utilized. And, LMV, LAvgPro, LMaxPro, LUV and LWV represent the respective aggregation approach in layered scheme. In case of multilayer approach, three classifiers as PART, RBFN, NBT are used in first layer and aggregator with two classifiers SMO and LR are used in second layer.

Data preprocessing is the first and most important step, so towards that data cleaning and transformation have applied as in Australian and Japanese datasets have some missing values and all datasets except than Taiwan & German-Numeric datasets are having some categorical and some numerical features. So, samples with missing sample are eliminated and label encoding has been utilized for transformation. Further, preprocessed datasets are segregated by 10-Fold-Cross-Validation. As some datasets are imbalance to a specific class, so here same number of sample of each class are kept in each fold and samples are randomly assigned to folds. Further, 10-Fold-Cross-Validation with 50 iterations have been utilized for comparative analysis.

Result as are tabulated in Table 3 on Taiwan dataset, results are mean of 10-Fold-Cross-Validation with 50 iterations. From the results, it is observed that PART has accomplished the utmost classification accuracy as compared to NBT, LRA, RBFN and SMO. Dagging with NBT, Rotation Forest with PART, Random Subspace with MLP, Decorator with LR and Multiboost with SMO have achieved best accuracies. In Table 5 and in Table 6, "No" represents results obtained by without applying any ensemble learning approaches with respective classification approaches in respective dataset. All aforementioned classifiers with various ensemble learning approaches have the significant improvement towards to classification accuracy. With various ensemble learning approach PART with Rotation Forest has the best classification performances as compared to nine ensemble learning with five classifiers. As compare to various ensemble classifier frameworks WV with single layer and multilayer approach have achieved better accuracies and also upgraded the classification accuracy as evaluated against to its base classifiers and other ensemble classifier frameworks. And, from the experimental observation including ensemble learning approaches with various classification and ensemble frameworks have also progressed the classification accuracy. Overall, Dagging and Multiboost with WV in layered approach have the best and second best classification accuracies. Overall, Dagging has the best classification accuracies almost with all classifiers and all ensemble frameworks have better improvements in classification performances.

Similar to Taiwan dataset, with other datasets namely: Bank-marketing in Table 4, German-categorical in Table 5, German-numerical in Table 6, Japanese in Table 7

Table 3 Classification accuracy of various classification approaches, classification approaches with various ensemble learning approaches and various ensemble classification frameworks with various ensemble learning approaches on Taiwan dataset

	No	Bagg	Cvpar	Adab	Deco	Subs	ROFo	Dagg	Metac	Multib
NBT	79.68	81.01	80.43	81.08	81.16	81.39	81.60	82.44	80.62	81.63
PART	81.43	82.49	82.25	81.46	81.54	81.88	82.87	82.67	82.61	82.18
MLP	79.10	79.96	79.89	80.54	80.62	80.79	80.33	79.67	81.80	80.07
LR	81.04	81.82	81.85	81.76	81.93	80.57	81.85	81.74	81.66	81.89
SMO	80.93	81.75	81.79	81.74	81.82	79.17	80.72	79.57	80.15	81.98
MV	82.68	82.75	82.71	82.61	82.77	82.81	82.97	83.11	82.88	83.01
AvgPro	82.57	83.19	82.60	82.70	81.83	81.98	82.87	83.18	82.85	82.71
MaxPro	82.60	82.89	82.65	82.78	81.85	82.03	82.89	83.10	82.74	82.61
UV	80.05	82.39	80.39	80.14	80.36	80.11	80.66	80.92	79.91	80.18
WV	82.69	82.84	82.77	82.70	82.26	82.98	83.06	83.20	82.97	83.10
LMV	83.59	83.98	83.67	83.43	83.00	83.64	83.80	83.94	83.71	83.84
LAvgPro	83.40	83.62	83.42	83.53	82.65	82.80	83.70	84.02	83.68	83.54
LMaxPro	83.43	83.31	83.48	83.61	82.67	82.85	83.72	83.93	83.57	83.43
LUV	80.85	83.21	81.17	80.94	81.16	80.91	81.47	81.73	80.71	80.99
LWV	83.76	83.83	83.76	83.69	83.25	83.98	84.06	84.20	83.97	84.10

Table 4 Classification accuracy of various classification approaches, classification approaches with various ensemble learning approaches and various ensemble classification frameworks with various ensemble learning approaches on Bank-marketing dataset

	No	Bagg	Cvpar	Adab	Deco	Subs	ROFo	Dagg	Metac	Multib
NBT	83.99	90.57	90.25	89.58	90.34	90.12	90.59	89.85	90.37	90.41
PART	88.30	90.17	89.18	89.76	90.23	89.74	90.72	90.39	89.87	90.19
MLP	88.59	89.76	89.54	89.20	90.37	89.41	89.38	89.41	89.52	89.83
LR	89.87	90.61	90.77	90.77	90.63	89.56	90.77	90.70	90.61	90.77
SMO	88.48	89.38	89.36	89.79	90.30	89.36	89.36	89.38	89.41	89.56
MV	90.37	90.14	90.46	89.96	90.43	89.49	91.27	89.74	89.87	90.57
AvgPro	90.30	90.92	90.37	89.83	90.54	89.36	90.87	90.76	89.83	91.21
MaxPro	89.36	89.67	89.36	89.58	89.41	89.36	90.26	89.72	89.43	89.99
UV	87.77	88.48	87.77	88.83	88.48	88.56	89.46	88.59	88.48	88.90
WV	90.46	91.23	90.55	91.05	91.52	90.58	91.36	91.83	89.96	91.66
LMV	91.45	91.62	91.54	91.54	91.52	91.57	92.37	92.82	91.95	92.65
LAvgPro	91.38	91.00	91.45	90.91	91.63	90.43	91.96	90.84	90.91	91.29
LMaxPro	90.43	90.75	90.43	90.66	90.48	90.43	91.35	90.79	90.50	91.07
LUV	88.65	89.36	88.65	89.72	89.36	89.45	90.35	89.47	89.36	89.79
LWV	91.46	91.23	91.55	91.05	91.53	91.58	92.38	92.83	91.96	92.86

Table 5 Classification accuracy of various classification approaches, classification approaches with various ensemble learning approaches and various ensemble classification frameworks with various ensemble learning approaches on German-categorical dataset

	No	Bagg	Cvpar	Adab	Deco	Subs	ROFo	Dagg	Metac	Multib
NBT	74.83	76.46	75.84	74.93	75.35	74.83	77.37	75.12	75.53	75.74
PART	71.66	76.05	72.32	73.83	73.73	74.34	76.86	74.34	73.02	75.65
MLP	71.91	75.95	72.09	74.03	73.61	77.47	76.05	76.75	73.73	75.14
LR	75.99	76.46	76.66	76.66	77.47	76.13	76.66	76.76	76.76	76.26
SMO	75.84	76.86	76.15	76.26	76.26	75.31	76.15	76.25	76.76	76.66
MV	77.16	77.67	76.66	77.25	77.05	77.62	77.76	77.35	77.95	77.26
AvgPro	77.06	77.47	75.95	77.74	77.06	77.82	77.97	78.56	77.15	78.49
MaxPro	75.95	77.37	75.95	75.74	76.14	76.11	77.06	77.85	77.35	77.56
UV	71.93	75.14	71.93	72.86	72.96	69.80	75.54	74.55	73.16	74.97
WV	77.94	78.45	77.43	76.00	76.81	73.35	77.53	78.10	76.71	78.02
LMV	78.57	79.22	78.19	76.75	79.21	78.07	79.30	79.85	78.47	79.78
LAvgPro	77.83	78.24	76.71	75.49	77.83	73.55	78.04	78.39	77.92	78.32
LMaxPro	76.79	78.22	76.79	76.56	76.97	76.91	77.91	78.43	77.17	78.40
LUV	72.64	75.89	72.64	73.59	73.69	70.49	76.29	75.29	73.89	75.72
LWV	78.71	79.23	78.20	78.76	78.58	78.08	78.92	79.98	79.48	79.91

and Australian in Table 8 are tabulated. With Bank-marketing dataset, MultiBoosting improves the classification performance of all approaches and with LWV has the best classification accuracy. In case of German categorical dataset, Bagging with LWV

Table 6 Classification accuracy of various classification approaches, classification approaches with various ensemble learning approaches and various ensemble classification frameworks with various ensemble learning approaches on German-numerical dataset

	No	Bagg	Cvpar	Adab	Deco	Subs	ROFo	Dagg	Metac	Multib
NBT	72.99	76.46	74.84	73.93	74.94	72.72	77.57	74.03	74.24	75.95
PART	70.62	76.05	72.32	73.83	72.62	74.34	77.16	75.14	72.22	76.15
MLP	71.89	76.86	73.93	75.45	73.63	71.61	74.84	73.23	71.81	74.64
LR	76.99	76.46	76.66	76.66	76.76	74.54	77.77	77.16	76.86	77.97
SMO	76.74	74.03	76.15	76.26	76.66	71.31	77.27	75.55	77.16	77.27
MV	77.32	77.42	77.52	75.99	78.44	73.44	78.74	77.40	76.70	77.52
AvgPro	76.81	78.23	76.36	76.75	77.16	72.11	77.87	76.95	77.67	77.77
MaxPro	76.70	77.83	75.21	75.73	76.15	76.11	77.82	77.94	78.17	78.28
UV	72.13	78.18	72.43	75.81	75.50	73.23	78.90	77.25	76.32	77.04
WV	77.47	77.57	77.68	76.14	78.59	76.59	78.90	78.55	76.86	78.68
LMV	78.09	78.19	78.30	77.75	79.22	78.17	79.53	79.16	78.47	79.30
LAvgPro	77.57	79.02	77.12	77.51	77.94	77.84	78.65	78.71	78.45	78.55
LMaxPro	77.62	78.76	77.06	77.61	77.07	77.98	78.40	78.84	79.11	79.21
LUV	73.57	75.74	73.62	74.32	74.01	74.70	74.48	74.80	74.85	74.56
LWV	78.25	78.35	78.45	76.90	79.38	74.32	79.69	79.32	78.63	79.45

Table 7 Classification accuracy of various classification approaches, classification approaches with various ensemble learning approaches and various ensemble classification frameworks with various ensemble learning approaches on Japanese dataset

	No	Bagg	Cvpar	Adab	Deco	Subs	ROFo	Dagg	Metac	Multib
NBT	85.03	87.24	84.90	86.95	85.34	85.34	87.24	86.51	85.34	87.89
PART	83.91	88.27	84.75	87.83	85.63	86.80	86.51	86.51	85.34	87.09
MLP	83.77	86.95	86.80	84.67	85.92	84.46	86.07	86.51	86.95	86.51
LR	85.94	86.51	85.48	87.21	87.26	85.63	86.51	85.92	85.48	86.51
SMO	85.65	86.66	86.51	87.79	87.23	83.43	86.66	87.37	87.52	87.08
MV	86.75	88.80	87.24	87.96	87.09	86.63	87.76	87.62	87.74	87.81
AvgPro	86.75	88.40	87.39	87.74	87.24	86.22	86.66	87.24	87.52	84.90
MaxPro	86.03	87.52	86.83	86.63	86.22	86.22	86.95	87.39	86.66	88.41
UV	83.16	86.08	84.07	83.08	85.45	85.57	86.66	86.30	85.35	85.75
WV	86.84	88.89	87.33	88.05	87.18	86.71	87.84	87.71	87.83	87.90
LMV	87.62	89.69	88.11	88.84	87.97	87.49	88.63	88.49	88.62	88.69
LAvgPro	87.62	89.28	87.47	88.62	88.11	87.08	87.52	88.11	88.39	85.75
LMaxPro	86.46	87.96	87.27	87.06	86.65	86.65	87.38	87.82	87.09	87.53
LUV	84.08	85.22	85.00	84.00	85.55	85.66	86.76	86.40	85.45	86.69
LWV	87.71	89.78	89.07	88.93	88.05	87.58	88.72	89.58	88.92	89.75

has the best classification accuracy. With German numerical dataset, Rotation Forest with LWV has accomplished the utmost classification accuracy. With Japanese dataset, Bagging with LWV has accomplished the utmost classification accuracy. With German

Table 8 Classification accuracy of various classification approaches, classification approaches with various ensemble learning approaches and various ensemble classification frameworks with various ensemble learning approaches on Australian dataset

	No	Bagg	Cvpar	Adab	Deco	Subs	ROFo	Dagg	Metac	Multib
NBT	82.75	85.88	84.43	86.46	85.45	85.88	85.59	87.33	84.58	86.90
PART	83.62	86.66	83.71	86.51	85.48	86.80	88.56	86.22	85.48	87.09
MLP	84.93	86.51	82.84	85.19	86.36	85.48	86.95	87.68	85.92	85.19
LR	86.96	85.63	87.04	87.83	87.53	87.24	87.68	87.39	86.80	87.83
SMO	85.51	86.95	85.59	85.92	86.36	85.19	86.36	87.68	86.36	85.92
MV	87.32	87.39	86.88	86.05	87.81	87.08	87.81	88.38	87.96	88.54
AvgPro	86.61	87.24	85.74	87.24	86.66	86.51	87.83	88.12	86.80	88.56
MaxPro	86.16	87.24	85.16	87.09	86.22	86.66	87.53	88.12	85.92	87.97
UV	84.64	84.96	83.71	84.53	84.15	83.86	84.77	85.54	83.85	84.91
WV	87.76	87.82	87.32	86.49	88.25	87.51	88.25	88.82	88.40	88.99
LMV	88.20	88.27	87.76	86.92	88.70	87.96	88.70	89.27	88.85	89.44
LAvgPro	87.39	88.03	86.51	88.03	87.43	87.29	88.62	88.91	87.58	89.36
LMaxPro	87.56	88.20	86.68	88.20	87.61	87.46	88.79	89.09	87.76	89.53
LUV	85.78	85.81	84.41	85.21	85.71	84.45	84.92	85.83	84.51	85.47
LWV	88.81	88.88	88.37	87.52	89.31	88.57	89.31	89.89	89.46	90.05

numerical dataset, MultiBoosting with LWV has accomplished the utmost classification accuracy.

From the experimental results as are depicted in Tables 3-8 with respective datasets, without applying any ensemble approaches LR have achieved better accuracies with most of the datasets, Rotation Forest, Bagging and Multiboost have accomplished the utmost classification accuracy with single layered approaches aggregated by WV respectively with respective datasets, Multiboost and Dagging with WV in layered approach is the best ways to improve the classification performances of credit scoring datasets. Overall, from the experimental observations on six credit scoring datasets, it can be concluded that Layered-WV approach with heterogeneous classifiers with MultiBoost as ensemble learning approach is the best way for credit scoring data classification.

5.3 Comparative Analysis with Prior Studies

This subsection presents a comparative analysis of the outcomes accomplished (specifically classification accuracy) from prior works and outcomes obtained from this study. Simulation results obtained by various approaches applied in this study with respective datasets as are tabulated in Tables 3-8, from these table the best accuracy achieved in respective dataset are as tabulated in Table 9. And, results obtained from the previous work as tabulated in Table 1. So, comparative graph of in between results from prior work as are tabulated in Table 1 and from this study as in Table 9 with respective datasets are depicted in Fig. 5. From the Fig. 5, it is visible that with Japanese dataset this study have achieved best, with Australian dataset fifth best and with German-categorical dataset sixth best accuracy.

As, the prior approaches applied for credit scoring are categorized into three categories as “classification”, “ensemble”, and “hybrid”. From the Table 1, it is observed that results

Table 9 Overall accuracies in credit scoring datasets

Dataset	Accuracy
Taiwan	84.20
Bank-marketing	92.86
German-categorical	79.91
German-numerical	79.69
Japanese	89.78
Australian	90.05

obtained by NRS base feature selection with layered ensemble framework has achieved best accuracy. And, by considering classification and ensemble approaches, Vertical Bagging with DT (VBDTM) has achieved the best accuracy. As, this study has presented the results analysis on ensemble learning and ensemble framework in previous sub-section. So, by comparing the results of classification and ensemble approaches from prior work and results from this study (learning approaches with multilayer ensemble classifier), results of this study have achieved second best accuracy. And, overall, this study has achieved fifth best performer, and NRS+LWV and GFSS are the best performer in Australian and German dataset respectively. But, out of best five, four approaches have applied feature selection approach and eliminated the redundant or irrelevant features from the datasets. So, by applying the feature selection approach with this study may also improve the classification performance.

6 Conclusion

Credit scoring is a prominent issue in the banking or financial sector, and slight improvement in its predictive performances would have a great impact. Various studies have shown that ensemble learning and ensemble framework are the approaches to get close to ideal classifier and it strengthens the classifiers by combining different models. But, from literature it not clear that which combination is the best way to improve the predictive performance. So, this study have presented a comparative analysis with nine ensemble learning approaches "such as Bagging, Cross Validation Parameter, Adaboost, Decorator, Subspace, Rotation Forest, Dagging, Metacoast, Multiboost" with various classification approaches such as PART, RBFN, LR, NBT and SMO along with various ensemble classifiers framework with layered and single layer with various aggregation approaches such as Majority Voting, Average Probability, Maximum Probability, Unanimous Voting and Weighted Voting. And, its impact on six benchmark credit scoring datasets "namely: Taiwan, Bank-marketing, German-categorical & numerical, Japanese and Australian" obtained from UCI Repository. From the experimental outcomes, it is observed that Multiboost and Dagging are best ensemble learning approaches and these approaches also improved the classification performances. Multilayer ensemble classifiers framework is the finest method to progress the predictive measures. Overall, MultiBoost and Dagging with multilayer ensemble frame is the best approach for credit score classification, and it also improved the significant performance of various classifiers as well ensemble learning approaches.

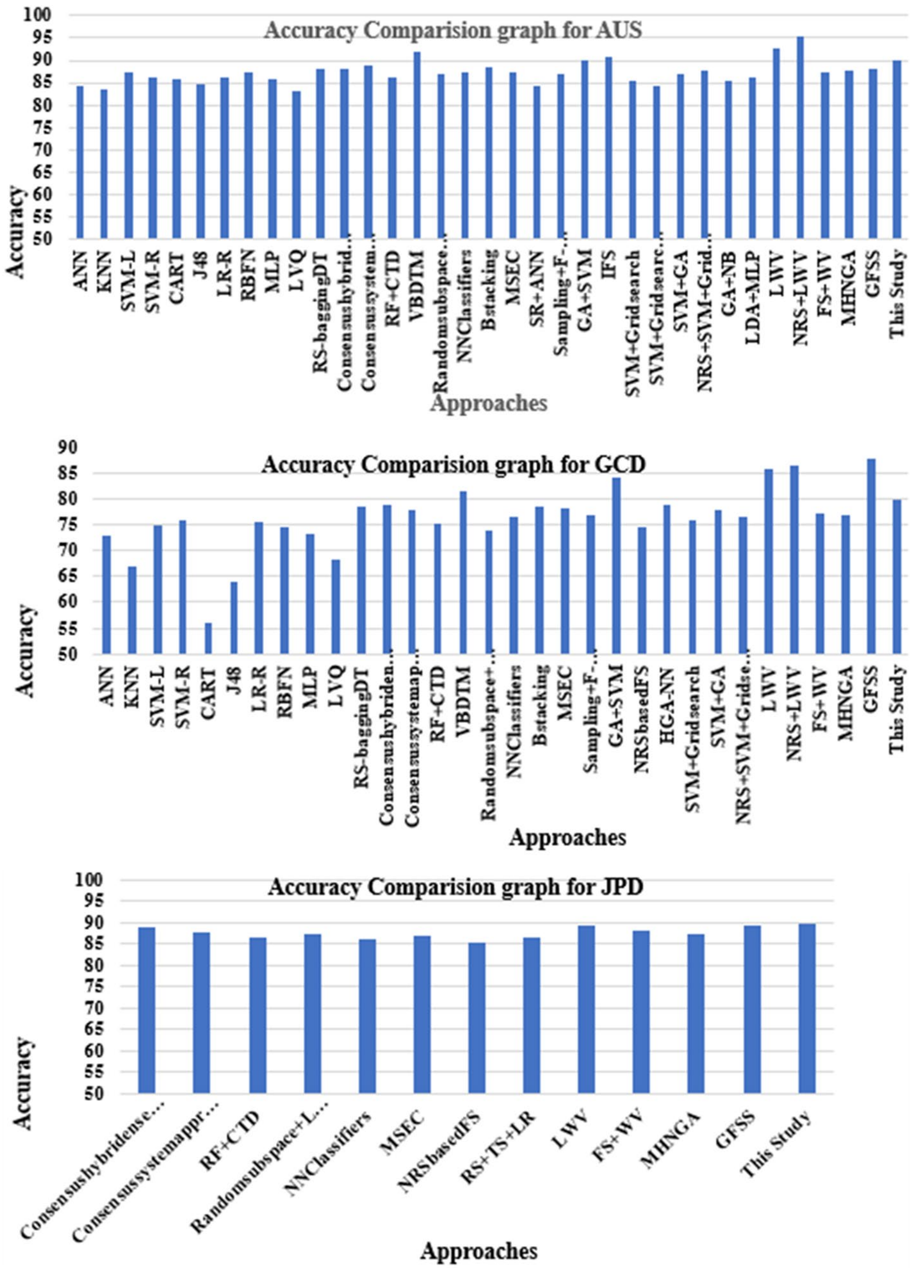

Fig. 5 Comparative graph of accuracies obtained from state-of-the-approaches and this study in Australian, German and Japanese datasets

Author Contributions A. K. Shukla, B. R. Reddy, G. S. Bopche, D. Chandramohan: These authors contributed equally to this work.

References

1. Mester, L. J., et al. (1997). What's the point of credit scoring? *Business review, 3,* 3–16.
2. Thomas, L.C., Edelman, D.B. & Crook, J.N. (2002). Credit scoring and its applications. *Journal of the Operational Research Society, 57*, 997–1006.
3. Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science, 21*(2), 117–134.
4. Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research, 201*(2), 490–499.
5. Kuppili, V., Tripathi, D. & Reddy Edla, D. (2020). Credit score classification using spiking extreme learning machine. *Computational Intelligence 36*(2), 402–426.
6. Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems, 26*, 61–68.
7. Sun, J., & Li, H. (2012). Financial distress prediction using support vector machines: Ensemble vs. individual. *Applied Soft Computing, 12*(8), 2254–2265.
8. Marqués, A., García, V., & Sánchez, J. S. (2012). Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications, 39*(12), 10916–10922.
9. Tripathi, D., Edla, D. R., & Cheruku, R. (2018). Hybrid credit scoring model using neighborhood rough set and multi-layer ensemble classification. *Journal of Intelligent & Fuzzy Systems, 34*(3), 1543–1549.
10. Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications, 73*, 1–10.
11. Parvin, H., MirnabiBaboli, M., & Alinejad-Rokny, H. (2015). Proposing a classifier ensemble framework based on classifier selection and decision tree. *Engineering Applications of Artificial Intelligence, 37*, 34–42.
12. Saha, M. (2019). Credit cards issued. http://www.thehindu.com/business/Industry/Credit-cards-issued-touch-24.5-million/article14378386.ece (2017 (accessed October 1)).
13. Vapnik, V. (2013). *The nature of statistical learning theory*. NY: Springer.
14. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273–297.
15. Van Gestel, T., et al. (2006). Bayesian kernel based classification for financial distress detection. *European journal of operational research, 172*(3), 979–1003.
16. Yang, Y. (2007). Adaptive credit scoring with kernel learning methods. *European Journal of Operational Research, 183*(3), 1521–1536.
17. Zhou, L., Lai, K. K., & Yen, J. (2009). Credit scoring models with auc maximization based on weighted svm. *International journal of information technology & decision making, 8*(04), 677–696.
18. XIAO, W.-b. & Fei, Q. (2006). A study of personal credit scoring models on support vector machine with optimal choice of kernel function parameters [j]. *Systems Engineering-Theory & Practice* **10**, 010.
19. Li, S.-T., Shiue, W., & Huang, M.-H. (2006). The evaluation of consumer loans using support vector machines. *Expert Systems with Applications, 30*(4), 772–782.
20. West, D. (2000). Neural network credit scoring models. *Computers & Operations Research, 27*(11), 1131–1152.
21. Haykin, S. S. (2001). *Neural networks: A comprehensive foundation*. NY: Tsinghua University Press.
22. Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on neural networks, 12*(4), 929–935.
23. Tripathi, D., Edla, D. R., Kuppili, V., & Bablani, A. (2020). Evolutionary extreme learning machine with novel activation function for credit scoring. *Engineering Applications of Artificial Intelligence, 96*, 103980.
24. Tripathi, D., Edla, D. R., Kuppili, V., & Dharavath, R. (2020). Binary bat algorithm and rbfn based hybrid credit scoring model. *Multimedia Tools and Applications, 79*(43), 31889–31912.
25. Tripathi, D. *et al.* Bat algorithm based feature selection: Application in credit scoring. *Journal of Intelligent & Fuzzy Systems* (Preprint), 1–10 .

26. Ala'raj, M., & Abbod, M. F. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems with Applications, 64,* 36–55.
27. Yeh, I.-C., & Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications, 36*(2), 2473–2480.
28. Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications, 38*(1), 223–230.
29. Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications, 36*(2), 3028–3033.
30. Zhang, D., Zhou, X., Leung, S. C., & Zheng, J. (2010). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications, 37*(12), 7838–7843.
31. Lin, W. .-Y., Hu, Y. .-H., & Tsai, C. .-F. (2012). Machine learning in financial crisis prediction: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(4), 421–436.
32. Lahsasna, A., Ainon, R. N., & Teh, Y. W. (2010). Credit scoring models using soft computing methods: A survey. *The International Arab Journal of Information Technology, 7*(2), 115–123.
33. Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management, 18*(2–3), 59–88.
34. Bequé, A.., & Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications, 86* 42–53.
35. Ala'raj, M., & Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems, 104,* 89–105.
36. Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications, 34*(4), 2639–2649.
37. Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications, 93*, 182–199.
38. Guo, S., He, H., & Huang, X. (2019). A multi-stage self-adaptive classifier ensemble model with application in credit scoring. *IEEE Access, 7*, 78549–78559.
39. Wongchinsri, P. & Kuratach, W. (2017). *Sr-based binary classification in credit scoring*, 385–388 (IEEE).
40. Hens, A. B., & Tiwari, M. K. (2012). Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method. *Expert Systems with Applications, 39*(8), 6774–6781.
41. Huang, C.-L., & Wang, C.-J. (2006). A ga-based feature selection and parameters optimizationfor support vector machines. *Expert Systems with applications, 31*(2), 231–240.
42. Hu, Q., Yu, D., Liu, J., & Wu, C. (2008). Neighborhood rough set based heterogeneous feature subset selection. *Information sciences, 178*(18), 3577–3594.
43. Liu, Y., et al. (2011). An improved particle swarm optimization for feature selection. *Journal of Bionic Engineering, 8*(2), 191–200.
44. Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert systems with applications, 41*(4), 2052–2064.
45. Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications, 33*(4), 847–856.
46. Ping, Y., & Yongheng, L. (2011). Neighborhood rough set and svm based hybrid credit scoring classifier. *Expert Systems with Applications, 38*(9), 11300–11304.
47. Liang, D., Tsai, C.-F., & Wu, H.-T. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems, 73*, 289–297.
48. Wang, J., Guo, K., & Wang, S. (2010). Rough set and tabu search based feature selection for credit scoring. *Procedia Computer Science, 1*(1), 2425–2432.
49. Edla, D. R., Tripathi, D., Cheruku, R., & Kuppili, V. (2018). An efficient multi-layer ensemble framework with bpsogsa-based feature selection for credit scoring data analysis. *Arabian Journal for Science and Engineering, 43*(12), 6909–6928.
50. Tripathi, D., Edla, D. R., Kuppili, V., Bablani, A., & Dharavath, R. (2018). Credit scoring model based on weighted voting and cluster based feature selection. *Procedia Computer Science, 132*, 22–31.
51. Zhang, W., He, H., & Zhang, S. (2019). A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring. *Expert Systems with Applications, 121*, 221–232.
52. Xu, D., Zhang, X., & Feng, H. (2019). Generalized fuzzy soft sets theory-based novel hybrid ensemble credit scoring model. *International Journal of Finance & Economics, 24*(2), 903–921.

53. Tripathi, D., Cheruku, R., & Bablani, A. (2018). *in Relative performance evaluation of ensemble classification with feature reduction in credit scoring datasets* (pp. 293–304). Ny: Springer.
54. Somol, P., Baesens, B., Pudil, P., & Vanthienen, J. (2005). Filter-versus wrapper-based feature selection for credit scoring. *International Journal of Intelligent Systems, 20*(10), 985–999.
55. Wang, D., Zhang, Z., Bai, R., & Mao, Y. (2018). A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring. *Journal of Computational and Applied Mathematics, 329*, 307–321.
56. Tripathi, D., Edla, D. R., Bablani, A., Shukla, A. K., & Reddy, B. R. (2021). Experimental analysis of machine learning methods for credit score classification. *Progress in Artificial Intelligence*, 1–27.
57. Frank, E. & Witten, I.H. (1998). *Generating accurate rule sets without global optimization*. University of Waikato: Department of Computer Science.
58. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
59. Kala, R., Vazirani, H., Khanwalkar, N., & Bhattacharya, M. (2010). Evolutionary radial basis function network for classificatory problems. *IJCSA, 7*(4), 34–49.
60. Broomhead, D. S., & Lowe, D. (1988). *Radial basis functions, multi-variable functional interpolation and adaptive networks*. Royal Signals and Radar Establishment Malvern (United Kingdom): Tech. Rep.
61. Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied statistics, 191–201*,
62. Green, S., & Salkind, N. (2010). *Using spss for windows and macintosh: Analyzing and understanding data*. Uppersaddle River: Prentice Hall Google Scholar.
63. Trevor, H., Robert, T. & JH, F. (2017). *The elements of statistical learning: data mining, inference, and prediction*. Springer open.
64. Rokach, L. & Maimon, O.Z. *Data mining with decision trees: theory and applications*, Vol. 69. World scientific.
65. Kohavi, R. (1996). *Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid.*, Vol. 96, 202–207 (Citeseer).
66. Rifkin, R.M. (2002). *Everything old is new again: a fresh look at historical approaches in machine learning*. Ph.D. thesis, MaSSachuSettS InStitute of Technology.
67. Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods, 3*, 185–208.
68. Brown, G. (2011). *in Ensemble learning 312–320*. Springer.
69. Woźniak, M., Graña, M., & Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion, 16*, 3–17.
70. Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*(1–2), 1–39.
71. Ravikumar, P. & Ravi, V. (2006). *Bankruptcy prediction in banks by an ensemble classifier*, 2032–2036 (IEEE).
72. Breiman, L. (1996). *Bagging predictors. Machine learning, 24*(2), 123–140.
73. Aslam, J. A., Popa, R. A., & Rivest, R. L. (2007). On estimating the size and confidence of a statistical audit. *EVT, 7*, 8.
74. Kohavi, R. (1995). *Wrappers for performance enhancement and oblivious decision graphs*. Tech. Rep.: Carnegie-Mellon Univ Pittsburgh Pa Dept of Computer Science.
75. Freund, Y., Schapire, R. E., et al. (1996). *Experiments with a new boosting algorithm* (Vol. 96, pp. 148–156). NY: Citeseer.
76. Melville, P., & Mooney, R. J. (2003). *Constructing diverse classifier ensembles using artificial training examples* (Vol. 3, pp. 505–510). NY: Citeseer.
77. Ho, T.K. (1995). Random decision forests, Vol. 1, 278–282 (IEEE).
78. Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence, 28*(10), 1619–1630.
79. Ting, K. M. & Witten, I.H. (1997). Stacking bagged and dagged models.
80. Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive, 155–164 (ACM).
81. Webb, G. I. (2000). Multiboosting: A technique for combining boosting and wagging. *Machine learning, 40*(2), 159–196.
82. Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning, 36*(1–2), 105–139.
83. Bashir, S., Qamar, U., & Khan, F. H. (2016). Intellihealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework. *Journal of biomedical informatics, 59*, 185–200.

84. Liang, D., Tsai, C.-F., Dai, A.-J., & Eberle, W. (2018). A novel classifier ensemble approach for financial distress prediction. *Knowledge and Information Systems, 54*(2), 437–462.
85. Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(3), 226–239.
86. Triantaphyllou, E. (2000). *in Multi-criteria decision making methods 5–21*. Springer.
87. Lichman, M. (2013). UCI machine learning repository. http://archive.ics.uci.edu/ml.
88. Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems, 62*, 22–31.
89. Statlog. (2019). German dataset. https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/ ((accessed October 1)).
90. Statlog. (2019). Australian credit approval data set. http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/australian/australian.dat ((accessed October 1)).
91. Dua, D. & Graff, C. (2017). UCI machine learning repository. http://archive.ics.uci.edu/ml.

Diwakar Tripathi received BE degree in computer science and engineering from Institution of Electronics and Telecommunication Engineering, New Delhi, India, in 2009, ME degree in computer engineering (software engineering) from the Institute of Engineering and Technology, Devi Ahilya Vishwavidyalaya, Indore, India, in 2014 and Ph. D. from National Institute of Technology Goa, India in 2018. Currently he is working as assistant professor at Thapar institute of Engineering and Technology Patiala, Punjab, India. His current research includes Machine Learning and Data Analytics.

Alok Kumar Shukla received B.Tech degree in Computer Science & Engineering from UPTU, Lucknow, India, in 2010, M.E degree in Information Technology (Information Security) from the Institute of Engineering and Technology, Devi Ahilya Vishwavidyalaya, Indore, India, in 2014 and Ph.D. from National Institute of Technology Raipur, India in 2019. Currently, He is working as Assistant Professor at GL Bajaj Institute of technology and Management Greater Noida, India. His research is centered in bioinformatics, network security, and machine learning domains.

B Ramachandra Reddy is working as a Senior Assistant Professor in the Department of Computer Science and Engineering at SRM University-AP Andhra Pradesh India, India. He received Ph.D. from PDPM IIITDMJ,Jabalpur. His research interests are Machine Learning, Data Mining, Software Metrics and Software Quality.

Ghanshyam S. Bopche is an Assistant Professor at the Department of Computer Applications, National Institute of Technology Tiruchirappalli (NITT) India. His research areas include Cyber Security, Digital Forensics and Technologies for Cyber Défense, Cloud Computing, and Machine Learning. He received his B.Sc. (Electronics) in 2007, MCA (Master of Computer Applications) in 2010 from the Nagpur University, and PhD (Cyber Security) from the IDRBT (Institute for Development and Research in Banking Technology, an associate institute of University of Hyderabad), India in 2017. He was exchange research scholar at the State University of New York (SUNY) at Buffalo, NY, USA during 2015.

D Chandramohan received the Ph.D. degree in computer science & engineering from Pondicherry Central University, Puducherry, India, and He is currently Associate Professor, Department of Computer Science and Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Andhara Pradesh, India. His area of interest includes Distributed Web Service, Web Service (Evaluation) Testbed, Software Metrics, GVANET and Cloud Computing, Opportunistic Computing, Evolutionary Computing, Service Computing, Software Engineering, Multi-Agent, Pervasive & Ubiquitous Computing, Fog & Edge Computing, Underwater Communication, Privacy and Security. Currently he is working on E-Waste Management, Disaster Management, Bio-Inspired Algorithms and Privacy Preserving Generic Framework for Cloud Data Storage, Optimization approach for minimizing Agro-crops. He is having 12-Years of academic and research expertise and 3-years of industrial experience.

Authors and Affiliations

Diwakar Tripathi[1] · Alok Kumar Shukla[2] · B. Ramachandra Reddy[3] · Ghanshyam S. Bopche[4] · D. Chandramohan[5]

Alok Kumar Shukla
alokjestshukla@gmail.com

B. Ramachandra Reddy
brreddy@iiitdmj.ac.in

Ghanshyam S. Bopche
ghanshyambopche.mca@gmail.com

D. Chandramohan
pdchandramohan@gmail.com

[1] Thapar Institute of Engineering and Technology, Patiala, Punjab 147004, India

[2] VIT University AP-Andhra Pradesh, Amaravati, Andhra Pradesh 522237, India

[3] SRM University AP - Andhra Pradesh, Amaravati, Andhra Pradesh 522502, India

[4] National Institute of Technology Tiruchirappalli, Tiruchirappalli, Tamilnadu 620015, India

[5] Madanapalle Institute of Technology and Science, Madanapalle, Andhra Pradesh 517325, India