Journal of Autonomous Intelligence (2023) Volume 6 Issue 2
doi: 10.32629/jai.v6i2.824

ORIGINAL RESEARCH ARTICLE

A novel credit scoring system in financial institutions using artificial intelligence technology

Geetha Manikanta Jakka[1], Amrutanshu Panigrahi[2], Abhilash Pati[2,*], Manmath Nath Das[3], Jyotsnarani Tripathy[4]

[1] *Department of Information Technology, University of Cumberlands, Williamsburg 40769, Kentucky*

[2] *Department of CSE, Siksha O Anusandhan University, Bhubaneswar 751030, India*

[3] *Department of AI & DS, VNR Vignana Jyothi Institute of Engineering &Technology, Hyderabad 500090, India*

[4] *Department of CSE-AIML & IoT, VNR Vignana Jyothi Institute of Engineering &Technology, Hyderabad 500090, India*

*** Corresponding author:** Abhilash Pati, er.abhilash.pati@gmail.com

ABSTRACT

In order to evaluate a person's or a company's creditworthiness, financial institutions must use credit scoring. Traditional credit scoring algorithms frequently rely on manual and rule-based methods, which can be tedious and inaccurate. Recent developments in artificial intelligence (AI) technology have opened up possibilities for creating more reliable and effective credit rating systems. The data are pre-processed, including scaling using the 0–1 normalization method and resolving missing values by imputation. Information gain (IG), gain ratio (GR), and chi-square are three feature selection methodologies covered in the study. While GR normalizes IG by dividing it by the total entropy of the feature, IG quantifies the reduction in total entropy by adding a new feature. Based on chi-squared statistics, the most vital traits are determined using chi-square. This research employs different ML models to develop a hybrid model for credit score prediction. The ML algorithms support vector machine (SVM), neural networks (NNs), decision trees (DTs), random forest (RF), and logistic regression (LR) classifiers are employed here for experiments along with IG, GR, and chi-square feature selection methodologies for credit prediction over Australian and German datasets. The study offers an understanding of the decision-making process for informative characteristics and the functionality of machine learning (ML) in credit prediction tasks. The empirical analysis shows that in the case of the German dataset, the DT with GR feature selection and hyperparameter optimization outperforms SVM and NN with an accuracy of 99.78%. For the Australian dataset, SVM with GR feature selection outperforms NN and DT with an accuracy of 99.98%.

Keywords: credit scoring system; machine learning (ML); classification techniques; feature selection algorithms; hyperparameter optimization

ARTICLE INFO

Received: 28 June 2023
Accepted: 20 July 2023
Available online: 22 August 2023

1. Introduction

A credit rating is a method used to assess the reliability of a potential lender, be it an individual or a company. The result of a credit score calculation is typically expressed as a numeric value, with a higher value indicating a more creditworthy potential lender. Those borrowers' evaluation scores will be higher and have profiles as close as possible to those of on-time debtors in the past. The bank decides whether to grant or deny the credit based on the criteria above by selecting a cut-off criterion[1]. The fundamental objective of the credit-scoring method is to build realistic models to aid in the lending industry's and banking industry's economic decision-making processes. Statistical and machine learning (ML) techniques use historical registration data to assess the potential risk posed by an application.

Credit scoring systems may suffer if the final data contains irrelevant or redundant information[2]. The financial industry has shown much interest in credit scoring because it is essential for managing credit risk. Machines, instead of humans, utilize artificial intelligence (AI). Using a method ensures that the correct conclusion will be reached whenever the technique is applied to a problem. Algorithms can also handle a scenario that is both complex and unpredictable. In contrast, conventional economic and financial theory uses the central von Neumann-Morgenstern axioms by employing very general reducing conventions. To better understand how commercial financing officers reach court rulings, we propose an AI algorithmic throughput model study algorithmic decision-making networks[3]. The throughput model is an AI computational model that encapsulates the concepts of evaluations, assessments, and choices.

Since even a small improvement in the credit scoring model can result in substantial earnings for banking organizations[4], many AI and ML models have been deployed to evaluate the efficacy of binary categorization in credit scoring. There has been substantial development in AI approaches for classification problems like credit rating. ML methods have been widely used in the credit rating industry in recent years. These classification problems have been tackled using SVM methods for ML[5,6], ANNs[7,8], and DTs[9,10]. Despite the advancements in ML methods, traditional statistical methods such as linear discriminant analysis (LDA) and logistic regression (LR)[7] continue to dominate the market[8] due to their ease of implementation.

Both businesses and universities have recognized the value of credit scoring for evaluating clients' financial stability. Credit scoring models are developed to aid financial institutions in making decisions regarding loan approval and credit line increases. With even a 1% improvement in the credit score model's classification reliability, financial institutions could drive more revenue while considerably reducing losses from potential bad debt. Over the past two decades, many different approaches to credit scoring have been created. Classification methods can be divided into three broad categories: standalone classifiers, uniform ensembles, and mixed-type ensembles. The results of using multiple classifiers in an ensemble rather than just one are more favorable[9]. The predictive power of the proposed method is put to the test in several different credit scoring models. The results of the experiments reveal that using these AI models for credit scoring in financial institutions greatly enhanced the anticipated achievement and that the final prediction accuracy of the suggested model is higher than that of other comparable models. This proves the usefulness of the presented model and opens up a new avenue of research into ML for the future.

1.1. Objective and contributions

The main objective of this research work is to employ different ML models to develop a hybrid model for credit score prediction. The ML algorithms support vector machine (SVM), neural networks (NNs), decision trees (DTs), random forest (RF), and logistic regression (LR) classifiers are employed here for experiments along with information gain (IG), gain ratio (GR), and chi-square feature selection methodologies for credit prediction over Australian and German datasets.

The contribution of this work can be summarized as follows:

- To develop a hybrid ML model for predicting the credit score;
- To implement information gain (IG), gain ratio (GR), and chi-square as the feature selection algorithm to deal with the noisy data;
- To implement SVM, NN, DT, RF, and RF as classifiers for credit score prediction;
- To evaluate the proposed over two datasets known as Australian and German datasets, with four evaluation parameters as accuracy, F-Measure, precision, and recall;

1.2. Paper structure

The structure of this article is summarized as follows. Section 1 shows the introduction along with the

contribution of this work. Section 2 represents the literature survey done for the work. Section 3 holds the problem statement of the work. Section 4 shows the proposed work's model, dataset description, and model development methodologies. The empirical analysis of the proposed work is done in section 5. Finally, section 6 represents this work's overall conclusion and future scope.

2. Related work

A new multi-stage ensemble credit scoring model has been proposed incorporating outlier adaption strategies to improve forecast accuracy[10]. The dataset is cleaned of missing values, its quantitative properties are standardized, and its categorical labels are transformed into dummy variables during preprocessing. The local outlier factor (LOF) method identifies and removes data outliers. This ensemble model includes a three-stage process. In the first stage, initial projections are provided by one of four classification algorithms: LR, DT, RF, or SVM. Next, the LOF technique identifies and eliminates any anomalies from the initial forecasts. The remaining projections are integrated using a weighted average ensemble technique in the third stage. This research evaluates the ensemble model compared to state-of-the-art techniques like support vector machines (SVMs) and extreme learning machines (ELMs). The results demonstrate that this ensemble method outperformed the rest regarding accuracy, AUC, sensitivity, and specificity. The performance of the proposed ensemble model is evaluated against several distinct scenarios in a sensitivity analysis. Results demonstrate that the ensemble model is robust to variations in input parameters and performs consistently better than competing techniques. The study introduces an innovative approach to credit scoring that uses outlier adaption techniques to improve the reliability of predictions. Financial organizations may find this ensemble model helpful when considering whether or not to extend loans. The model has to be tested on more datasets, and its potential applications should be explored in greater depth. In particular, the research[11] focuses on the popular ensemble models used in credit evaluation. However, sophisticated tree-based classifications are rarely used in ensemble models. Only a small number of researchers have considered dynamic ensemble selection. In an effort to rectify the current literature gap, this paper will offer a novel tree-based over-fitting cautious heterogeneous ensemble model (i.e., OCHE) for credit scoring. Prediction accuracy and processing costs for base models are both optimized by tree-based methods. The proposed method could dynamically apply base weights to models based on the over-fitting metric for ensemble selection. Several state-of-the-art methods, including neural networks (NNs), gradient boosting, DTs, and RFs, were put through their paces alongside the researchers' proposed method. The results of this investigation showed that the ensemble method's created model outperformed any alternative tactics. In terms of computing cost, the proposed solution can be considerably improved using GPU acceleration. The system's potential applications need to be explored, and it has to be evaluated on other datasets, but this requires further research. Credit ratings could be efficiently categorized using the spiking extreme learning machine (SELM), as described in the study of Kuppili et al.[12]. The leaky nonlinear integrate and fire model (LNIF) proposes a novel function for producing spikes. The ELM calculates the time between spikes and uses that information to group credit ratings. The SELM framework is tested on several credit score databases, including those for Australia, Germany (categorical and numeric), Japan, and bankruptcy. SELM results are also compared to those obtained by backpropagation, probabilistic NNs, ELMs, voting-based Q-generalized ELMs, radial basis neural networks, and ELM utilizing certain current spiking neuron models in terms of classification efficacy, area under the curve (AUC), h-measure, and computational time. According to the findings, the suggested SELM significantly improves accuracy and execution time for the databases above. Therefore, the efficiency of ELM's categorization is enhanced by adding a biological spiking function. The spiking function of the body will be used to improve the efficiency and precision of credit score classification. Financial institutions use credit scoring to evaluate loan default risk. However, due to a dearth of credit statistics, P2P lending is limited in generating reliable credit scores. Credit ratings use various alternative data sources to make up for the scarcity of financial data. Much interest

has been shown in the study's primary research[13]: financial institutions' access to social network data can improve their prediction powers. This research aims to determine how accurately social network data can foretell loan default. Debtors' social media data was extracted from their mobile devices, and then LR was used to analyze the correlation between that information and loan defaults. Three artificial intelligence algorithms (RF, AdaBoost, and LightGBM) were created to demonstrate the precision of social network data prediction. LR's findings support statistical analysis of social network data and loan default. But other information about social networks, including how often people talk on the phone, is not collected. This limits research into how social networks might be used to assess credit risk.

3. Problem statement

The problem statement for credit scoring can be summarized as the need for a reliable method to classify credit applicants according to their creditworthiness. Traditional ML techniques for credit rating may not be immune to the presence of outliers and have poor accuracy and processing efficiency[14]. Therefore, in order to address these concerns and improve the accuracy of credit scores, new approaches are needed. Articles under review provide state-of-the-art methods, such as SELM and a multi-stage ensemble model with outlier adaption techniques, to improve the accuracy of credit score forecasts. Testing these strategies on real-world datasets has shown that they outperform conventional ML strategies. More research is needed to evaluate the practicality of these algorithms and test them on larger datasets.

4. Proposed credit scoring system

This method is used to refine the classification process and boost its efficacy. The foundation of multiple classifier systems combines several classifiers to achieve better results than any of the individual classifiers. Most methods for developing classifiers revolve around modifying the training dataset, developing classifiers on these n new training sets, and merging the results into a single decision rule. **Figure 1** depicts the process flow of the suggested hybrid credit scoring system's design.

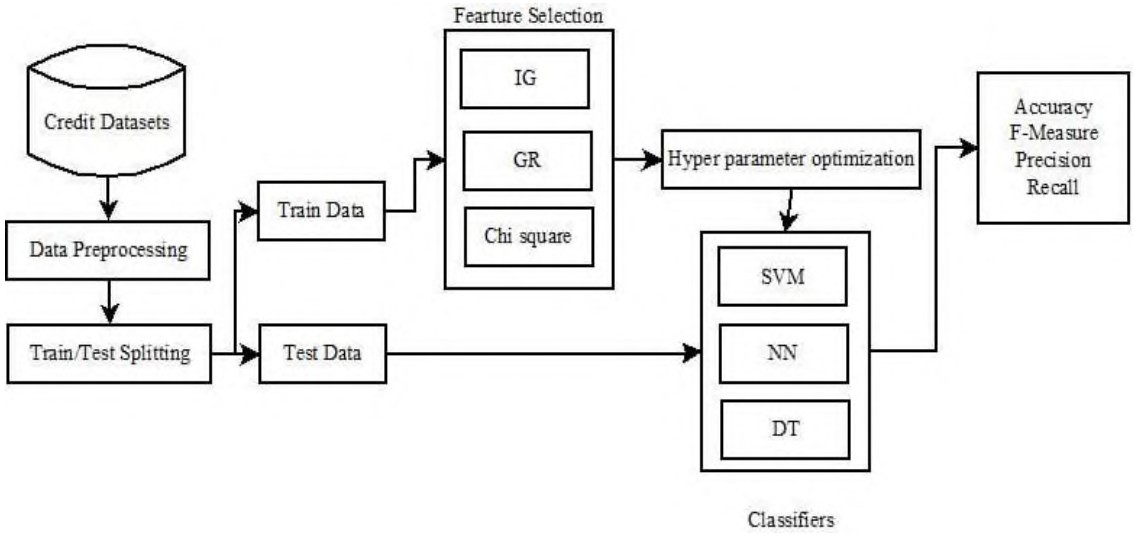

Figure 1. Flow diagram of novel credit scoring system.

4.1. Credit datasets

The study uses two credit datasets from conventional financial institutions or peer-to-peer lending. The UCI-ML repository makes all of the conventional credit datasets, which are well-liked and often used by study participants, accessible to the general public. The suggested model was validated using datasets from Australia and Germany; a detailed description is provided in **Table 1**[15].

Table 1. Considered datasets description.

Dataset	Attributes	Loans		
		Bad	Good	Total
German	20	300	700	1000
Australian	14	383	307	690

4.2. Partitioning and pre-processing of data

Dataset attributes are given distinct values to eliminate duplicate information and streamline the modeling procedure. Data transformation is necessary when working with information with a non-standard value scale. The attribute in the data set is normalized so that it can only take on values between 0 and 1. To accomplish this, we first find the greatest value for each property and then divide all the values for that characteristic by that number. Credit scoring algorithms in Germany consider a single database of credit scoring information available to the public. There are twenty-one variables in all, each representing a different component of the respondents in the German credit data (one of which is an outcome variable). This dataset contains information on a total of one thousand people who have made prior claims of authorship. Deflators and non-defaulters are the creditor's primary focus, with a "1" indicating "they are not defaulters" in 70% of cases and a "0" indicating "defaulter" in 30% of cases. Applications for new credit are evaluated using historical debtor information and other input factors. Credit records almost always need to have numerical values associated with them. As a result, data pre-processing is required before scoring algorithms can be developed to account for missing or slightly inaccurate variables. The missing-valued samples can be easily removed. However, when information is assumed to be randomly missing, missing data removal is effective. This assumption must be changed from a credit rating perspective. The charging technique is useful for dealing with missing data for non-random reasons, as it uses estimates to replace missing values. Next, they take the average value of the continuous features still in the data set and use that to fill in the gaps left by the missing data. As such, the 0–1 normalization method was used to scale the data in this investigation. Given a characteristic c, the normalized c' can be found in the formula below:

$$c' = \frac{c - \min(c)}{\max(c) - \min(c)'} \quad (1)$$

K-fold cross-validation is used to increase the algorithm's resilience. The initial data set is split into k sections of equal size. One of the k subsets is kept as a sample to assess the model, whereas the other k subsets are used to train the model. Each subgroup is a test set once throughout the procedure, which continues k times. One estimate for the entire dataset may be created by averaging the k evaluation results from the folds. Using either 5 or 10 is preferable since these numbers can result in a sensible trade-off between dependability and computing expense. Therefore, in this study, we carry out a five-fold cross-validation technique. To be more precise, run each dataset through a 5-fold cross-validation and execute procedure loops 50 times. The final result for each dataset is then produced by averaging the outcomes of this method[15].

4.3. Feature selection algorithms

Following, feature selection techniques are used. As not all 20 input features must be informative to predict the output classes in the case of the German credit dataset. Methods for feature selection are used to obtain this most useful feature collection. All three feature selection methods used to study—such as IG, GR, and chi-square—have been explained in this section[16].

4.3.1. Information gain (info-gain or IG)

It determines the information gained from the given being provided and gives individuals features to AI to make credit decisions and forecast the class accurately with the next example. These classes are determined

by comparing the trends of input predictors from trained AI models. The amount of information gained is determined by calculating the decrease in total entropy following adding a new feature. The anticipated value of a specific attribute needed for classifying an instance is called entropy. Assume that c and d are two variables, with c serving as an input feature for d's output. D's entropy is calculated as follows the equation:

$$H(D) = -\sum_{d \in D} K'(d) \log_2 K'(d) \tag{2}$$

The equation below shows an increase in entropy with the introduction of input predictor c.

$$H\left(\frac{D}{C}\right) = -\sum_{c \in C}(c) \sum_{d \in D} K'\left(\frac{d}{c}\right) \log_2 K'\left(\frac{d}{c}\right) \tag{3}$$

IG is the difference between the entropy of the predicting d before and after adding the input predictor c.

$$IG = H(D) - H(\frac{D}{C}) \tag{4}$$

$$IG = H(C) - H(\frac{C}{D}) \tag{5}$$

$$G = H(D) + H(C) - H(D, C) \tag{6}$$

IG is a proportionate, equal statistic whose value for d following the observation of c is equal to that for c following the observation of d.

4.3.2. Gain ratio (GR)

The GR is an expansion of IG. Especially when there is less information, IG exhibits a bias toward selecting the characters with more significant numbers. This demonstrates IG's flaw. The GR is an expanded version of IG. Especially when there is less information, IG is biased toward selecting characters with more significant numbers. This demonstrates IG's flaw.

$$GR = \frac{InfrG}{H(C)} \tag{7}$$

Whenever parameter d is predicted using the equation above, IG is normalized by dividing the total entropy of c. It provides a value for GR between 0 and 1 owing to normalization. If it is 1, the data in c will predict d; if it is 0, c and d will not be related to one another. Due to its preference for characteristics with smaller numerical values, the GR differs from IG.

4.3.3. Chi-square

Chi-square is used to identify a copy's most vital points. The chi-squared statistics are thoroughly examined, and it delivers valued characteristics from the characteristic space for the given class. This procedure tests the first hypothesis with the premise that the two traits are distinct.

$$\chi^2 = \sum_{q-1}^{j} \sum_{r-1}^{x} (\frac{K^{qr} - U^{qr}}{U^{qr}}) \tag{8}$$

where K refers to the frequency actually occurring, U refers to the anticipated frequency. A higher value of c denotes strong evidence for the first premise.

4.3.4. Hyperparameter optimization

Fine-tune the hyperparameters of the chosen model to optimize its performance. Utilizing techniques like grid search, random search, or Bayesian optimization to find the optimal set of hyperparameters that maximize accuracy. Extensive searching of a predetermined hyper-parameter space is conducted as part of the common hyper-parameter optimization technique known as grid search. The dimensionality scream afflicts grid search, however. Grid searches become significantly more expensive to compute as the number of hyper-parameters or the size of searching space rises. As a result, grid search is not appropriate for classifiers that have a large

number of hyper-parameters. Bayesian hyper-parameters optimization strategy because of its better effectiveness and speed. This method creates a statistical model that associates the hyper-parameters with a likely aim (often a cross-validated performance). The method then chooses hyper-parameters iteratively and evaluates their efficacy. In order to deduce as much information about the ideal hyper-parameters as possible, these data are accumulated, and a model is created. The Tree-Structured Parzen Estimator (TPE), a particular Bayesian hyper-parameters optimization method, is used in this research.

The hyper-parameters are regarded to be autonomous by TPE. Then provide the relevant object as w and the hyper-parameters G. TPE models $L_g(w \mid \lambda)$ for a statistical model G indirectly from $L_g(\lambda \mid w)$ and $L_g(w)$. The alternate density estimates in the TPE models $L_g(w \mid \lambda)$ are dependent on the value of c concerning a threshold w:

$$L_g(w|\lambda) = \begin{cases} \wp(\lambda), w < w^* \\ g(\lambda), w \geq w^* \end{cases} \quad (9)$$

where $\wp(\lambda)$ is a measure of density derived from measurements of the objective's worth, and it's less than w. The best-observed w is the quantile of w, with $\gamma = 0.15$ being the most frequent value. The anticipated improvement (EI) criteria, which chooses the present best-case hyper-parameters based on the available data, is used by TPE to determine the optimum. That the given expression, which consists of, γ $\wp(\lambda)$, and $g(\lambda)$:

$$EI(\lambda)\alpha\left(\gamma + \frac{g(\lambda)}{\wp(\lambda)}(1-\gamma)\right)^{-1} \quad (10)$$

This shows that EI may be maximized for hyper-parameter values by a high likelihood under $\wp(\lambda)$ and a low probability under $g(\lambda)$. To estimate $\wp(\lambda)$ and $g(\lambda)$, a Parzen estimator, also known as a Parzen-window estimator, is used. A 1D Parzen estimator is constructed to calculate each hyper-parameter chance density function. Since hyper-parameters are mutually independent, the combined density function $\wp(\lambda)$ or $g(\lambda)$ may be calculated by multiplying separate density calculations. Bayesian hyper-parameter optimization technique then iteratively chooses the most advantageous hyper-parameters and assesses them until it reaches a certain number of iterations. Compared to grid search, the Bayesian hyper-parameter optimization strategy produces better results. When choosing the appropriate split, the maximum number of characteristics determines how many features should be considered. Instead of the default "auto," which determines the hyper-parameter, this value is calculated employing the TPE algorithm.

$$\max fature = \sqrt{n\_feature} \quad (11)$$

where n features is the total number of collection features[15].

4.4. Classification technologies employed

As was already indicated, this research compares and uses three classifiers. The following subsections briefly explain SVM, NN, and DT techniques[17–19].

4.4.1. Support vector machine (SVM)

SVM is a classification method that has demonstrated its effectiveness as an AI method in various disciplines, including categorizing texts, credit risk, and predicting bankruptcy. SVM bases its modeling of a given system on structural risk minimization (SRM). An SVM uses structural risk minimization instead of statistical risk minimizing utilized by traditional NNs. Input vectors are nonlinearly mapped into feature space with high dimensions using SVMs, which then employ a linear model to contrivance nonlinear class boundaries. The greatest margin hyperplane in this high-dimensional space is discovered to maximize the distance between decision classes. The training samples that are closest to the maximum margin hyperplane are referred to as SVM. The approach is gaining popularity due to its numerous beneficial attributes and encouraging empirical results. SVM is an optimization method that concurrently reduces errors in prediction

and model complexity. Its main strength is this method's capacity to represent variability and produce intricate mathematical models. Using SVMs, the ideal hyperplane is found in new, maximizes high-dimensional space, the distance between it and closest training samples, and minimizes predicted generalization error.

4.4.2. Neural networks (NNs)

Massively parallel computers, known as NNs, tend to maintain experimental information and permit its future use. Inter-neural interconnections are employed to store the data, replicating the human brain intending to gather empirical evidence while learning. Concerning learning capabilities, NNs also can generalize newly acquired information. Several NN architectures and learning algorithms are in use today, as well as numerous applications for them. NNs are generally employed in economic situations where the variables are connected in non-linear ways. An ANN is a collection of neural nodes connected to weighted nodes. Every node may act as a creature's neuron, and the connections between the nodes are equivalent to the synapses that link the neurons. Input, hidden, and output layers are the three layers that make up the most popular kind of NN. Multilayered perceptron, or MLP, is the name of it. There is a connection between a layer of input units and a layer of hidden units, which connects layer output devices.

4.4.3. Decision tree (DT)

An instructed, acyclic network in the shape of a tree, a DT is an analysis of the data which encapsulates the probability of the class label about its predictive characteristics. There are no inbound edges at the DT's root. Each additional node has zero or other outgoing edges and precisely one incoming edge. A node is called a leaf node if it has no outward edges; otherwise, it is called an inner node. One class label is assigned to each leaf node, and one predictor attribute, the splitting characteristic, is assigned to each internal node. Every edge e coming from a node within n has an assumption q attached to it, and this predicate solely takes into account the splitting characteristic of n. Given the predictive attributes, a DT may be used to forecast the values of the goal or class attribute. To calculate the anticipated worth of an unidentified instance. Choose whether to enter the left or right child node depending on the value of the splitting property. Until you reach a terminal or leaf node, repeat this procedure using the splitting property for each new child node. The expected value for the target property is represented by the value of the target attribute shown in the leaf node. A DT may also be transformed into rules that might be applied to prediction jobs like bankruptcy and credit default.

4.4.4. Random forest (RF)

The RF is a well-known machine learning method that has found application in many different fields, including credit scoring. Credit scoring is the practice of assigning a numerical value between 300 and 850 that is supposed to indicate how likely a borrower is to be able to keep up with their monthly credit card or loan payments. Labeled data is necessary for RF's model training process as a supervised learning algorithm.

4.4.5. Logistic regression (LR)

The ML technique of logistic regression is also commonly employed in the credit rating process. It is a type of statistical analysis used to predict a binary outcome (such as "default" or "non-default") from a given collection of input features.

5. Results and discussions

The preceding four credit datasets from the actual world are utilized to confirm benchmark models and the suggested model using four assessment measures. On a system with a 3.0 GHz Intel i5 CPU, 8 GB RAM, and Microsoft Windows 10 OS, all experiments are programmed in Python 2.7. The German dataset has 1000 instances, 700 of which received credit, and 300 did not. The Australian dataset has 307 good and 383 bad loan instance data. Each of these cases has 20 judgment qualities, 7 of which are numerical and 13 of which are categorical, in the case of the German dataset. Any classification model's effectiveness accuracy (Acc) is

considered using a confusion matrix. This is the ratio of all occurrences that were correctly categorized to all other instances. It is efficient to determine classification accuracy using the F-Measure (F-M). The value is computed using precision and recalls harmonic mean. Following prediction, it assesses the fraction examples from the testing set that are defaulters and those that are not. A quick model is created using these measures. The duration of classifier training, expressed in seconds, is calculated using a stopwatch. Equations (12)–(15) show the performance parameters used to evaluate the proposed work[20,21].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{F} - \text{Measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{15}$$

When training and testing are split 70%–30%, accuracy, and F-Measure results are substantially equal for all classifiers while testing. The predictive value of these metrics should be strong. Each feature selection method was examined separately, and the results were recorded. The F-Measures and accuracy of several ML classifiers are shown in **Table 2** and **Figure 2**.

Table 2. F-Measure and accuracy for various ML classifiers.

Classifiers	Chi-square (in %)				GR (in %)				IG (in %)			
	German dataset		Australian dataset		German dataset		Australian dataset		German dataset		Australian dataset	
	Acc	F-M	Acc	F-M	Acc	F-M	Acc	F-M	Acc	F-M	Acc	F-M
SVM	90.01	92.86	91.01	93.17	90.11	92.92	91.88	93.93	90.89	93.58	93.04	94.89
NN	90.39	93.14	91.59	93.61	90.51	93.21	92.46	94.36	91.29	93.86	93.62	95.33
DT	92.21	93.01	91.31	93.39	90.29	93.07	92.17	94.14	91.09	93.72	93.33	95.12
RF	89.62	92.57	90.43	92.73	89.71	92.64	91.29	93.49	90.49	93.31	92.46	94.48
LR	89.83	92.71	90.72	92.95	89.91	92.78	91.59	93.71	90.71	93.44	92.75	94.69

The table mentioned above demonstrates that accuracy is a widely used statistic to assess how accurate the model's predictions are on the whole. It is determined as the proportion of occurrences that are appropriately categorized into all instances. On the other hand, F-Measure is another evaluation metric that considers both precision and recall. It balances these two metrics and is particularly useful when the classes are imbalanced. It can be observed from **Table 2** and **Figure 2** that NN surpasses almost all the cases excluding chi-square in the case of the German dataset, where DT outperforms all others in accuracy. Higher chi-square values indicate a stronger relationship between the feature and the target variable. For example, the chi-square value for DT is reported as 92.21% in the case of the German dataset. In comparison, the chi-square value for NN is reported as 91.59% in the case of the Australian dataset, indicating a strong relationship between the features used by the model and the target variable. Higher GR values indicate more informative features. For example, the GR value for NN is reported as 90.51% and 92.46% in the case of the German and Australian datasets, respectively, indicating that the features used by the NN model provide a high amount of information for classification. Higher IG values indicate more informative features. For example, the GR value for NN is reported as 91.29% and 93.62% in the case of the German and Australian datasets, respectively, indicating that the features used by the NN model provide a significant reduction in entropy and are highly informative for classification. The classifiers' F-Measure and accuracy metrics, achieved using the five-fold cross-validation

method, are shown in **Table 3** and **Figure 3**.

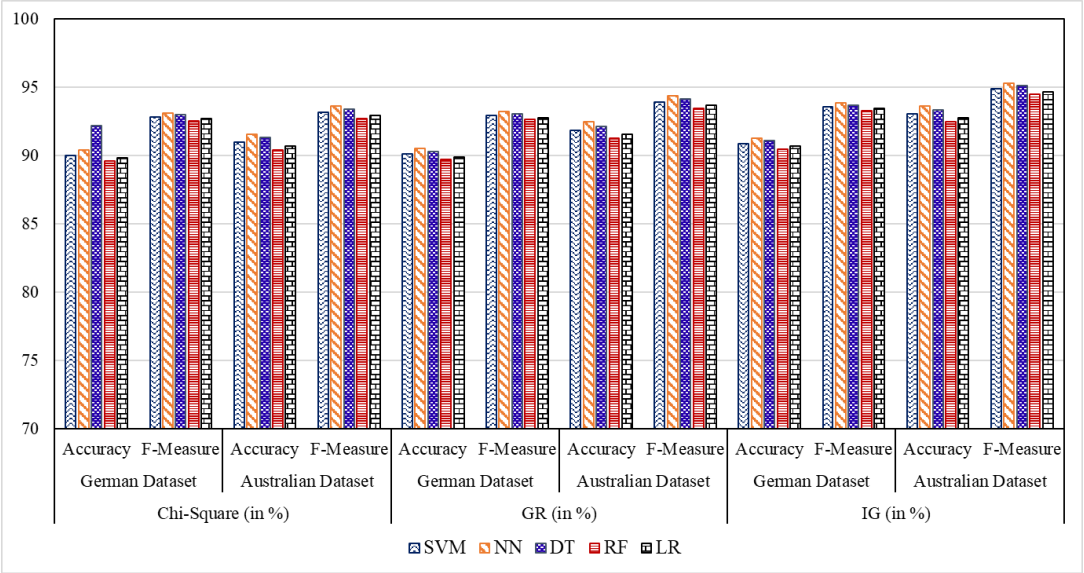

Figure 2. Obtained accuracy and F-Measure in % for various ML classifiers.

Table 3. F-Measure and accuracy of classifiers (5-Fold).

Classifiers	**Chi-square (in %)**				**GR (in %)**				**IG (in %)**			
	German dataset		**Australian dataset**		**German dataset**		**Australian dataset**		**German dataset**		**Australian dataset**	
	Acc	**F-M**	**Acc**	**F-M**	**Acc**	**F-M**	**Acc**	**F-M**	**Acc**	**F-M**	**Acc**	**F-M**
SVM	92.81	94.89	93.91	95.37	92.11	94.35	94.78	96.11	92.91	94.99	95.94	97.03
NN	92.41	94.57	94.49	95.81	92.51	94.64	95.36	96.53	93.31	95.27	96.52	97.53
DT	92.21	94.43	94.19	95.59	92.31	94.49	95.07	96.31	93.09	95.13	96.23	97.24
RF	91.59	94.01	93.33	94.93	91.71	94.07	94.19	95.66	92.49	94.71	95.36	96.59
LR	91.81	94.14	93.62	95.15	91.89	94.21	94.49	95.88	92.71	94.85	95.65	96.82

The table mentioned above presents the 5-fold cross-validated performance of SVM, NN, DT, RF, and LR classifiers. Accuracy and F-Measure are reported as average values, indicating the classifiers' overall correctness and balanced performance. Additional feature selection criteria, including chi-square, GR, and info-gain, demonstrate each classifier's relevance and informativeness. It can be observed that NN surpasses all others while applying 5-fold cross-validation to the previous outcomes with a very identical response in accuracies and F-Measures as well.

Table 3 examines the amount of time needed to train various classifiers and filtering methods for dividing data by 70%–30% and 5-fold cross-validation. As previously said, classifier training should take the least amount of time. **Figure 4** shows the training duration in seconds for the SVM, NN, DT, RF, and LR classifiers. The supplied numbers show how long each classifier needed to train on the provided dataset. Lower numbers denote a quicker training pace. The extra columns, chi-square, GR, and IG, probably show how long each classifier needed to use its own algorithms to complete feature selection.

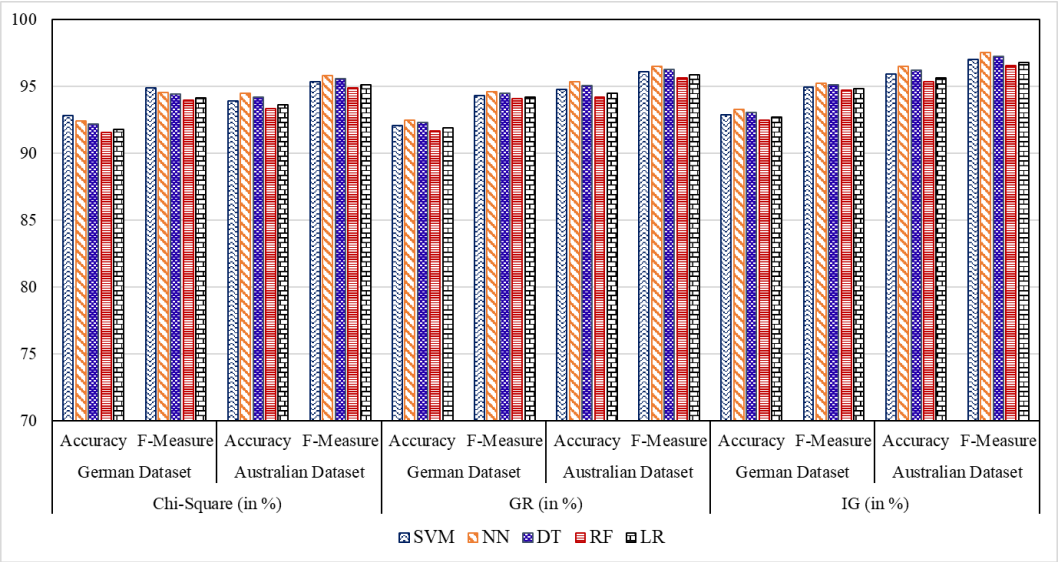

Figure 3. Obtained accuracy and F-Measure in % for various ML classifiers (5-fold).

Table 4. Training time (in Sec) of the ML classifiers.

Classifiers	For German dataset			For Australian dataset		
	Chi-square	GR	IG	Chi-square	GR	IG
SVM	3.01	0.01	0.01	2.64	0.01	0.02
NN	3.64	0.02	0.01	3.02	0.01	0.02
DT	3.43	0.29	0.33	2.88	0.21	0.28
RF	3.62	0.22	0.27	3.16	0.18	0.24
LR	3.18	0.18	0.24	3.01	0.16	0.22

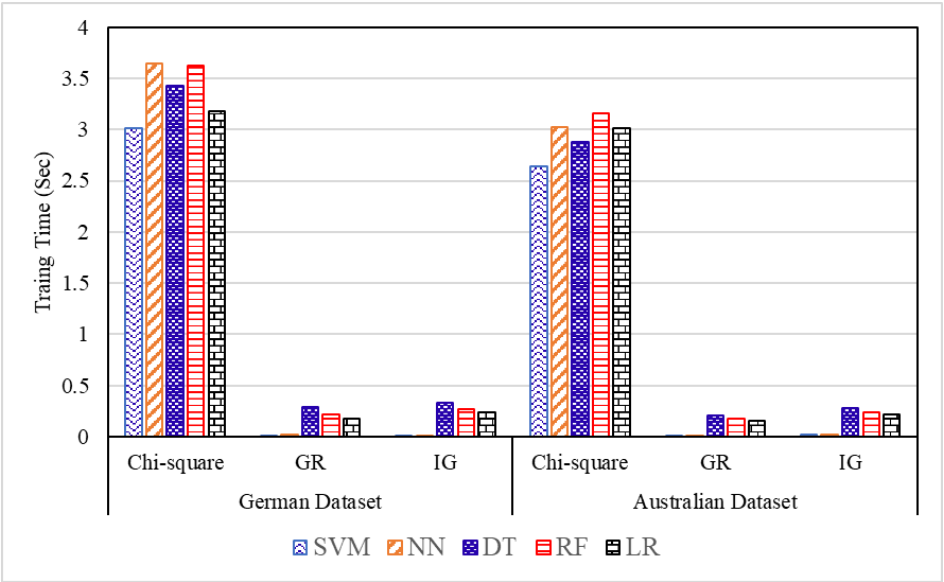

Figure 4. ML Classifiers' training time in seconds.

The evaluation metrics for GR as the feature selection technique with hyperparameter optimization are shown in **Table 5** for the German and Australian datasets and the SVM, NN, DT, RF, and LR classifiers. Metrics such as recall, precision, F-Measure, and accuracy are used for assessment. For each classifier, the values show how well it performed when tested with the specified dataset and measure. The classifiers effectively classify tasks on both datasets, achieving excellent performance and accuracy. **Figure 5** shows that DT achieved the highest accuracy (99.78%) on the German dataset, beating out the competition. SVM had the

highest F-Measure accuracy (99.89%), followed by DT (99.67%). Precision for NN was 99.76%, and recall for DT was 99.39%, both of which are the maximum achievable. However, SVM's accuracy (99.98%) on the Australian dataset was the highest of any method. The highest results for F-Measure and precision from DT were 99.89% and 99.87%, respectively, while the best result for recall from SVM was 99.67%.

Table 5. Performance analysis of various classifiers with hyper-parameter optimization.

Dataset	Evaluation measure	ML classifier (measured in %)				
		SVM	NN	DT	RF	LR
German	Accuracy	98.08	97.98	99.78	97.26	98.11
	F-Measure	99.89	98.78	99.67	98.88	99.22
	Precision	99.44	99.76	99.43	99.56	99.62
	Recall	98.67	99.34	99.39	98.21	98.82
Australian	Accuracy	99.98	98.78	97.53	97.81	98.46
	F-Measure	98.54	97.45	99.89	98.22	99.36
	Precision	97.67	98.43	99.87	98.56	99.62
	Recall	97.67	99.29	99.10	97.88	99.11

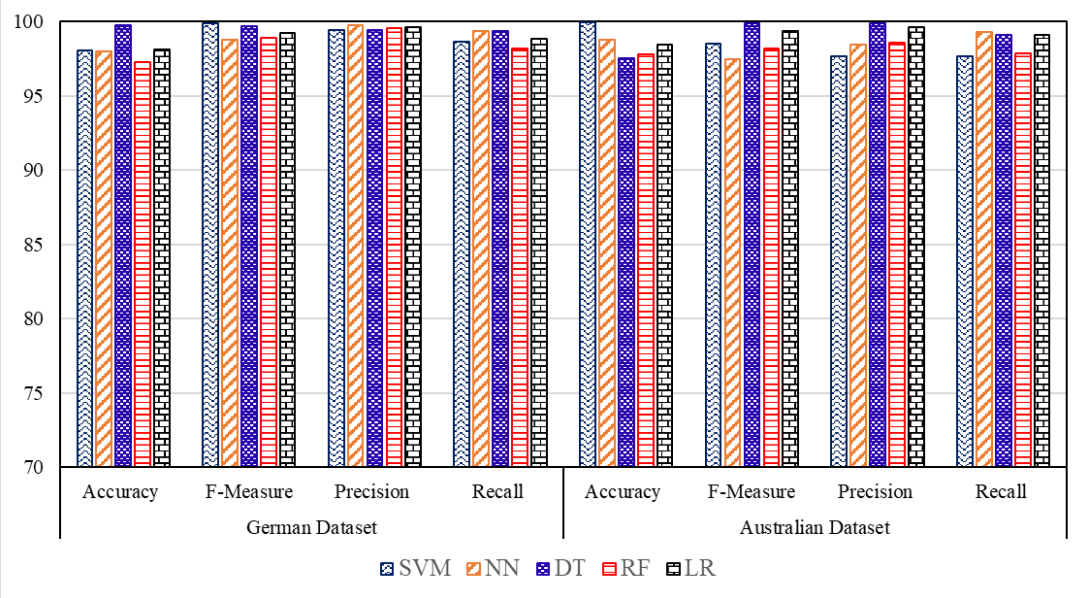

Figure 5. Performance comparison among different classifiers with hyper-parameter optimization.

The hybrid model employing ML and feature selection techniques offers significant advantages in credit score prediction. Still, they also come with challenges that need to be carefully addressed to ensure responsible and effective use in lending and financial decision-making. Balancing the benefits of automation and accuracy with interpretability, fairness, and regulatory compliance is crucial in deploying ML-based credit scoring systems.

The advantages of these proposed hybrid models can be listed as follows:

1) Improved predictive accuracy: From experiments, it can be noted that achieving enhanced predictive accuracies, i.e., more than 99% in both datasets considered.

2) Handling non-linearity: In this proposed hybrid model, the considered ML methods, such as NNs and SVMs, can capture complex non-linear relationships between credit-related variables that traditional methods might miss.

3) Feature selection and optimization: Several feature selection and optimization techniques are included here to make the trained model efficient.

4) Potential for real-time applications: This hybrid model can be deployed in predicting real-time, enabling faster decision-making processes in credit evaluation.

The disadvantages of these proposed hybrid models can be listed as follows:

1) Data requirement: ML algorithms often require a large amount of data for effective training, which might be challenging for organizations with limited historical credit data. Here the datasets used have 1000 and 690 instances only.

2) Computational resources: Training sophisticated ML models and performing feature selection on large datasets can be computationally intensive and may require substantial computing resources.

3) Domain expertise and interpretation: While ML models can automatically learn from data, they might not consider domain-specific knowledge that credit analysts possess, leading to potential oversights or inadequate consideration of contextual factors.

6. Conclusion and future scope

Finding credit people who default and for whom accurate data for forecasting is required requires using credit scoring. By contrasting three feature selection methods as well as five basic classifiers, this study was able to accomplish its objectives effectively. Additionally, two fusion procedures were utilized, one of which was a novel way described in this study, and the other was a standard method, that has been used in previous works. The findings of this study's credit-scoring measurements are consistent with the idea that integrating ML models with feature selection techniques can increase the overall accuracy of credit-scoring applications from a few percent to several percent. The empirical analysis shows that in the case of the German dataset, the DT with GR feature selection and hyperparameter optimization outperforms SVM, NN, RF, and LR by approximately 1.73%, 1.83%, 2.59%, and 1.70%, respectively, in terms of accuracy. For the Australian dataset, SVM with GR feature selection outperforms NN and DT by approximately 1.21%, 2.51%, 2.21%, and 1.54%, respectively, in terms of accuracy. This integration of ML approaches, along with feature selection techniques and hyperparameter tuning considering training-testing split with five-fold cross-validation, achieving a good predictive outcome, makes it a novel hybrid approach.

Further, the proposed model can be tested over a more significant time series dataset to verify the robustness of the model. In addition to enhancing the performance of the reported model, different ML-based optimization techniques can also be employed. We are planning for further works with new generation boosting algorithms (XgBoost, CatBoost, and Lightboost) and other ensemble classifiers on other Credit datasets available.

Author contributions

Conceptualization, GMJ; methodology, GMJ, AP (Amrutanshu Panigrahi) and AP (Abhilash Pati); software, AP (Amrutanshu Panigrahi); validation, AP (Abhilash Pati), MND and JT; formal analysis, GMJ; investigation, GMJ; resources, MND; data curation, JT; writing—original draft preparation, GMJ; writing—review and editing, AP (Amrutanshu Panigrahi) and AP (Abhilash Pati); visualization, AP (Amrutanshu Panigrahi); supervision, AP (Abhilash Pati).

Ethics approval and consent to participate

Not applicable.

Acknowledgment

The authors would like to thank themselves.

Funding

This research received no external funding.

Conflict of interest

The authors declare no conflict of interest.

References

1. Pławiak P, Abdar M, Acharya UR. Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. *Applied Soft Computing* 2019; 84: 105740. doi: 10.1016/j.asoc.2019.105740
2. Tripathi D, Edla DR, Cheruku R, Kuppili V. A novel hybrid credit scoring model based on ensemble feature selection and multilayer ensemble classification. *Computational Intelligence* 2019; 35(2): 371–394. doi: 10.1111/coin.12200
3. Rodgers W, Hudson R, Economou F. Modeling credit and investment decisions based on AI algorithmic behavioral pathways. *Technological Forecasting and Social Change* 2023; 191: 122471. doi: 10.1016/j.techfore.2023.122471
4. Alaei F, Alaei A, Pal U, Blumenstein M. A comparative study of different texture features for document image retrieval. *Expert Systems with Applications* 2019; 121: 97–114. doi: 10.1016/j.eswa.2018.12.007
5. Ping Y, Yongheng L. Neighborhood rough set and SVM based hybrid credit scoring classifier. *Expert Systems with Applications* 2011; 38(9): 11300–11304. doi: 10.1016/j.eswa.2011.02.179
6. Zhang D, Zhou X, Leung SCH, Zheng J. Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications* 2010; 37(12): 7838–7843. doi: 10.1016/j.eswa.2010.04.054
7. Dumitrescu E, Hué S, Hurlin C, Tokpavi S. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research* 2022; 297(3): 1178–1192. doi: 10.1016/j.ejor.2021.06.053
8. Wei S, Yang D, Zhang W, Zhang S. A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning. *IEEE Access* 2019; 7: 99217–99230. doi: 10.1109/ACCESS.2019.2930332
9. Feng X, Xiao Z, Zhong B, et al. Dynamic ensemble classification for credit scoring using soft probability. *Applied Soft Computing* 2018; 65: 139–151. doi: 10.1016/j.asoc.2018.01.021
10. Zhang W, Yang D, Zhang S, et al. A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring. *Expert Systems with Applications* 2021; 165(4): 113872. doi: 10.1016/j.eswa.2020.113872
11. Xia Y, Zhao J, He L, et al. A novel tree-based dynamic heterogeneous ensemble method for credit scoring. *Expert Systems with Applications* 2020; 159: 113615. doi: 10.1016/j.eswa.2020.113615
12. Kuppili V, Tripathi D, Edla DR. Credit score classification using spiking extreme learning machine. *Computational Intelligence* 2020; 36(2): 402–426. doi: 10.1111/coin.12242
13. Niu B, Ren J, Li X. Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending. *Information* 2019; 10(12): 397. doi: 10.3390/info10120397
14. Faisal MF, Saqlain MNU, Bhuiyan MAS, et al. Credit approval system using machine learning: Challenges and future directions. In: 2021 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA); 9–10 December 2021; Tirana, Albania. pp. 76–82.
15. Xia Y, Liu C, Da B, Xie F. A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications* 2018; 93: 182–199. doi: 10.1016/j.eswa.2017.10.022
16. Trivedi SK. A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society* 2020; 63: 101413. doi: 10.1016/j.techsoc.2020.101413
17. Pati A, Parhi M, Pattanayak BK. A review on prediction of diabetes using machine learning and data mining classification techniques. *International Journal of Biomedical Engineering and Technology* 2023; 41(1): 83–109. doi: 10.1504/IJBET.2023.128514
18. Pati A, Parhi M, Pattanayak BK. IHDPM: An integrated heart disease prediction model for heart disease prediction. *International Journal of Medical Engineering and Informatics* 2022; 14(6): 564–577. doi: 10.1504/IJMEI.2022.126526
19. Ghodselahi A, Amirmadhi A. Application of artificial intelligence techniques for credit risk evaluation. *International Journal of Modeling and Optimization* 2011; 1(3): 243–249. doi: 10.7763/IJMO.2011.V1.43
20. Rout SK, Sahu B, Panigrahi A, et al. Early detection of sepsis using LSTM neural network with electronic health record. In: Ambient Intelligence in Health Care: Proceedings of ICAIHC 2022; 15–16 April 2022; Bhubaneswar, India. Springer; 2022. pp. 201–207.
21. Sahu B, Panigrahi A, Rout SK, Pati A. Hybrid multiple filter embedded political optimizer for feature selection. In:

2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP); 21–23 July 2022; Hyderabad, India. pp. 1–6.