International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Broad Review on Different Imbalanced Dataset Classification Approaches

Mr. K. Suresh Babu[1], Dr. B. Vara Prasada Rao[2], Mr. Y. Narasimha Rao[3*], Dr. J. Hari Kiran[4], Dr. Sai Chandana Bolem[5], Dr. Srinivasa Rao Battula[6], Dr. Koteswara Rao Chittepureddi[7], Mr. Y. Anil Kumar[8], Mr. Ratna Babu Chekka[9]

Submitted: 07/05/2023 **Revised**: 13/07/2023 **Accepted**: 05/08/2023

Abstract: The series problem in extensive data application is managing the imbalanced data. Hence, the imbalanced data classification system was introduced to collect the imbalanced data at the maximum possible rate. Several neural mechanisms have been designed to classify imbalanced data with a high accuracy rate. However, the complexity of the data makes the classification process difficult by increasing the computation cost, resource usage, and algorithm complexity. Hence, this review has detailed several classification model performances in different imbalanced databases. Finally, the performance analysis has been done to analyse the classification performance of each model. Hence, the robustness has been estimated based on precision, specificity, accuracy, and sensitivity. In addition, the merits and limitations of each model are also discussed in detail. Subsequently, based on the demerits, the classification models provided future directions to improve the imbalance data.

Keywords*: *Imbalanced data, data mining, deep learning, classifiers, over and under sampling, optimization algorithms

1. Introduction

Class imbalance categorization is used in data mining and ML when one or multiple classes are underserved in the dataset [1]. Numerous real-world categorization jobs exhibit class unbalance, which is a significant hurdle for the data mining community [2]. The primary challenge with these issues is the skewed distribution that impairs the effectiveness of traditional classification methods, as typical learning algorithms assume a training database, making it more challenging to forecast minority class cases [3, 4]. In recent years, numerous efforts have been made to address binary imbalanced class concerns, which comprise only two classes. However, multi-class imbalanced classification is utilized in various fields, including text categorization, identification of human influence, and diagnosis [5]. Regrettably, it could be incorrect to directly transfer the solutions provided for dual imbalanced class issues to multi-class expansion in quantity. It could now utilize specific algorithms to address multi-class difficulties [6]. Fortunately, the academic community has developed decomposition algorithms to address the problem of multi-class categorization [7]. The multi-class categorization issues are reduced into two class subtasks in this step taken, which are significantly easier to differentiate. Two well-known techniques are one against one and one against all [45]. Because one against all has established a synthetic class imbalance [8]. Moreover, it is not recommended to resolve issues with skewed starting distributions [9]. The latest research demonstrates that when working with a set curriculum and data-level issues, combining a multi-class breakdown including one classifier improves considerably [10].

These approaches have several intriguing applications in the case of unbalanced datasets. The scientific community has recently placed a premium on multi-class unbalanced data categorization [11, 12]. However, most available techniques for dealing with skewed classes are oversampling-based. Under sampling is quite beneficial in binary unstable situations, where it may overcome a number of the drawbacks of resampling [13]. There is indeed a shortage of specialized under-sampling techniques that account for the presence of many categories and can infer their relationships [14]. Multi-class unbalanced problems are substantially more challenging to solve due to the many types that must study and the numerous interactions between these classes

[1]*Associate Professor, Department of Computer Science and Engineering, Narasaraopet Engineering College, Narasaraopet, AP. India*
[2]*Professor, Department of Computer Science and Engineering, RVR&JC College of Engineering, Guntur, Andhra Pradesh, India.*
[3*]*Professor, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India.*
[4]*Associate Professor, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India.*
[5]*Assistant Professor, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India.*
[6]*Assistant Professor, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India.*
[7]*Assistant Professor, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India.*
[8]*Research Scholar, VIT-AP University, Amaravati, Vijayawada, India*
[9]*Associate Professor, Department of Computer Science and Engineering, RVR&JC College of Engineering,Guntur, Andhra Pradesh, India.*
** Corresponding Author Email: y.narasimharao@vitap.ac.in*

[15].Moreover, the Conventional solutions developed for two-class issues may become unfeasible or underperform due to their inability to simulate this difficult challenge [16]. There are presently few specialized multi-class techniques, and further testing is required in this field. Static- Synthetic-Minority Over-sampling-Technique (SMOTE) used an M-step resampling technique; here, M denotes the classes count [16, 17]. The resampling technique picks the class with the smallest size for each cycle and replicates a couple of examples of that original dataset class [18]. A common pattern in this field emphasizes the critical importance of considering the unique characteristics of SMOTE classes. Their training difficulty has been raised while functioning the multi-class imbalanced oversampling data and recommends a data-driven generalizing that can be integrated into any statistics multi-class remedy [19].

Recent oversampling techniques concentrate on exploiting data from many classes simultaneously and minimizing the effect of overlapped and noisy elements [20]. Prototype selection in a sampling procedure aims to decrease the classifier's comparison set to expand efficiency and reduce storage needs. Nevertheless, in an unbalanced circumstance, the aim changes, as the data balance during the distribution process is more important [21-22]. The support vectors have focused on producing a meaningful under-sampled dataset using a genetic algorithm-guided search [23]. First, many randomly under-sampled info subsets are made and then developed until they could enhance the arguably best-underrepresented dataset further regarding fitness value [24]. Similarly, in any evolutionary process, the scheme in which genomes represent solutions is critical. Hence, a broad review has been conducted on imbalanced data classification to identify each model's essential functions and drawbacks.

2. Imbalanced Dataset Classification

In the ML community, imbalance class is a hot topic of research. However, prior and contemporary research has demonstrated that class duplication has a more detrimental effect on the effectiveness of classifiers. This article discusses in length and objectively evaluates class overlapping from the perspective of unbalanced data and its impact on the classifier. Steps in imbalanced data classification are described in Figure 1.

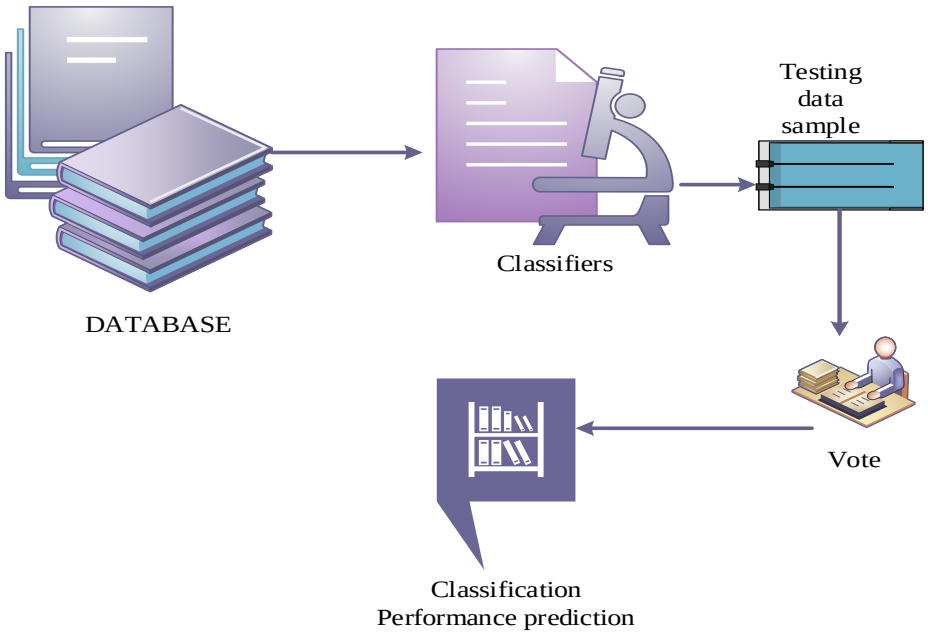

Fig 1: Steps of imbalanced data classification

First, we conduct a comprehensive experimental evaluation of class overlaps and imbalance. In contrast to past research, the experiment was conducted across the entire spectrum of class overlapping and an extensive range of imbalanced class degrees. Secondly, performed an in-depth technical analysis with existing methodologies for dealing with imbalanced datasets [25]. Current methods are analyzed critically and classified into class sharing and class coincides strategies. Additionally, emerging methods and recent developments in this field are examined in detail. Moreover, the experimental data in this research are similar to previous work in that they demonstrate unequivocally that the training algorithm's efficiency degrades with increasing degrees of category overlap. In contrast, imbalanced data does not even have an impact. The review highlights the importance of additional research into class overlap. Dang et al., [26] have made two significant contributions: first, a tuned deep networks for sewer fault diagnosis that is premised on a box architectural style and includes a sequence of fully

connected layers capable of efficiently extracting complex patterns from defective zones; and second, combination attachments of the developed framework that employ both an outfit strategy and an expense learning-based technique to address the highly imbalanced issue. The experimental findings demonstrated that the suggested architecture outperformed earlier sewer flaw sensing devices and was resistant to imbalanced data issues. The proposed defect prediction framework can encourage more effective defects given in equations and facilitate the degree of association of deep neural-based techniques in real-world sewage fault analysis applications

2.1 Oversampling in imbalanced data classification

Historically, this detection has been primarily accomplished through proper examination and oversight of cassava plantation by growers or extension services laborers from the agriculture management, and then noted to Agricultural-Advisory-services for further assessment. However, it is time-consuming capital costly. It cannot identify cassava disease in time to assist the farmers in applying preventative strategies to non-infected leaves and seeds to raise yields, increase Africa's market, and combat hunger. In addition, food marketing databases were utilized to check the robustness of the convolutional neural nets (CN), which is one of the imbalanced datasets [27]. Finally, it has earned the disease affection prediction exactness as 93%. An ensemble technique dubbed regression-vector-based voting-classifier (RVVC) has been designed for detecting hazardous remarks on social sites. Under benign voting conditions, the ensemble combines support vector and logic regression classifiers. Numerous tests are conducted on the unbalanced and balanced datasets to assess the operating performance of the proposed model. In addition, using the unbalanced dataset, an artificial minority oversampling method is utilized to restore data balance. Additionally, two extraction methods, maximum - likelihood, bag-of-words, and document frequency, is used to assess their appropriateness. The suggested approach's result is evaluated to that of many machine classifiers based on recall, accuracy, F-score, and precision. The recorded accuracy for oversampling is 0.95 and loss is 0.11; the loss gained for undersampling is 0.26, and the accuracy 0.88. In addition, without the Sampling model, the regression scheme has earned an accuracy of 0.93 and a loss of 0.13 [28].

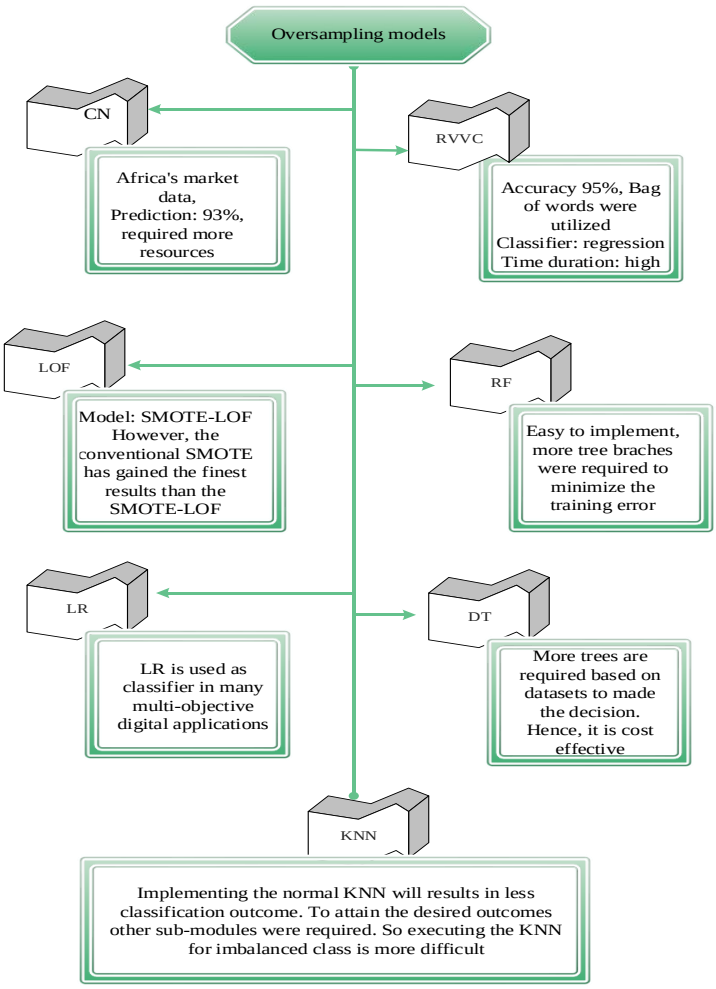

Fig 2: Remarks of discussed literatures

As a result, this work intends to enhance SMOTE's ability to detect noise in the synthetic data generated during the treatment of unbalanced data using the Local-Outlier-Factor (LOF). Hence, the model LOF was introduced by Maulidevi and Surendro [29] to gain the most acceptable classification outcome using the unbalanced data. Moreover, the results indicate that the designed SMOTE-LOF achieves higher precision and f-measurement than conventional SMOTE. However, for datasets with fewer test data samples, the AUC value for SMOTE was probably superior at managing skewed data. Random forests (RF) are the collection of decision trees in which the variable determines every tree is dependently taken separately and uniformly across the forest. In addition, if the trees counts were increased, then the generalization error coheres to the limit. Moreover, the generalization error in the tree classifiers of RF is proportional to the intensity of every tree in the RF.

The connection between one binary classifier and a group of alternative (explanatory) factors is investigated using logistic regression (LR) analysis. In addition, When the response variable has just dual input elements such as 0 and 1. Hence, this input element is used to answer the logistic questions in the form of yes or no. Moreover, the multimodal LR has more input elements to specify the multi-objective classes in a single run. Also, it is employed in several digital applications for detection purposes. The decision tree (DT) is a recurrent division of the data space that expresses a classifier. Furthermore, the clustering algorithm comprises vertices that create a grounded tree, which is a driven tree without incoming edges. Each of the remaining nodes has a single incoming edge. Moreover, an internal is a node that has outgoing advantages. Every additional node is referred to as leaves. Also, every inner leaf node divides the occurrence area into dual or multiple sub-spaces based on the discrete process of the input feature values. The k-Nearest-Neighbors (kNN) algorithm is a straightforward yet powerful classification technique. The primary disadvantage of k - means is its poor efficiency when utilized in dynamic applications. In addition, to gain the finest prediction or classification results, it depends on other sub-learning models. However, this KNN model is less complex and easy to implement. The remarks on discussed literature are explained in Figure 2.

2.2 Deep learning

Deep convolutional-based neural models (DCNM) were extensively investigated for the identification of skin diseases. Hence, some methods are achieving the finest diagnostic outcomes that are equivalent or superior to dermatologists. In addition, the widespread use of DCNM in the identification of skin diseases is hampered by the limited information and database imbalance of publicly available skin-lesion datasets. So, Yao et al., [30] have presented a unique technique for classifying skin lesions using a particular model on limited and unbalanced datasets. To begin, multiple DCNMs are learned on various small and unbalanced datasets to demonstrate that models of intermediate complexity outperformed models of greater complexity. Secondly, regularisation Drop Block and Drop Out are introduced to mitigate over fitting, and a Revised Rand Augment approach is provided to address the sampling underrepresentation problems in the short dataset. The accuracy of different deep neural architectures is described in Figure 3, and the classifier's performance is detailed in Figure 4.

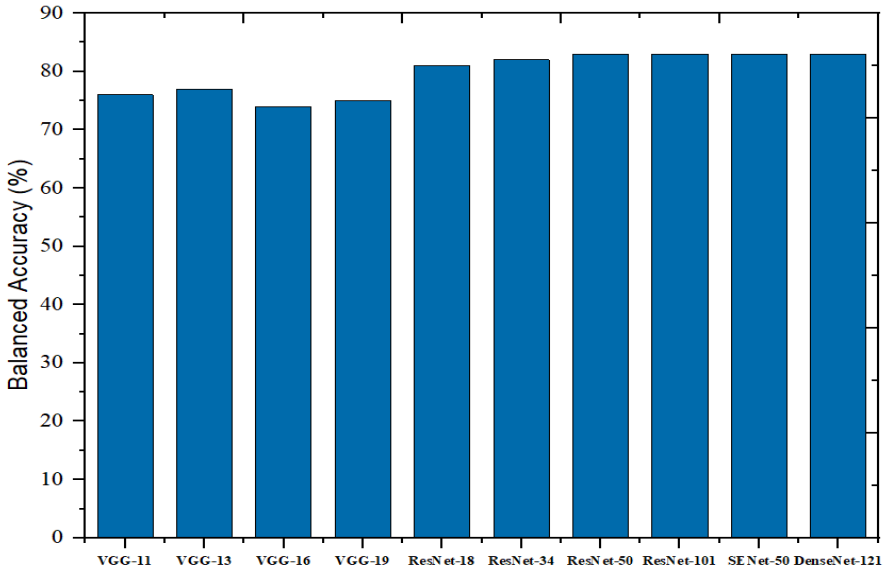

Fig 3: Accuracy validations of different deep neural architectures

Strip steel's surface flaws region is tiny, with various defect forms and complicated grey structures. Due to the high prevalence of false faults and interference from edge light, existing machine vision methods cannot identify faults in different kinds of strips of steel. Deep learning-based image identification approaches require many photos to train the network. However, the popular deep networks training activities can be accomplished on a database with a smaller sample with unbalanced class faults. Hence, Wan et al., [31] have provided a set of techniques for full-strip steel flaw identification based on quick pre-process algorithms and transfer learning mechanisms. These approaches have enabled quick screening of the surface, extraction of features from defects, balancing the categories in a sampling dataset, defeat prediction, data augmentation, and categorization. Moreover, the recognition accuracy gained by the upgraded VGG19 network was determined to be 97.8%.

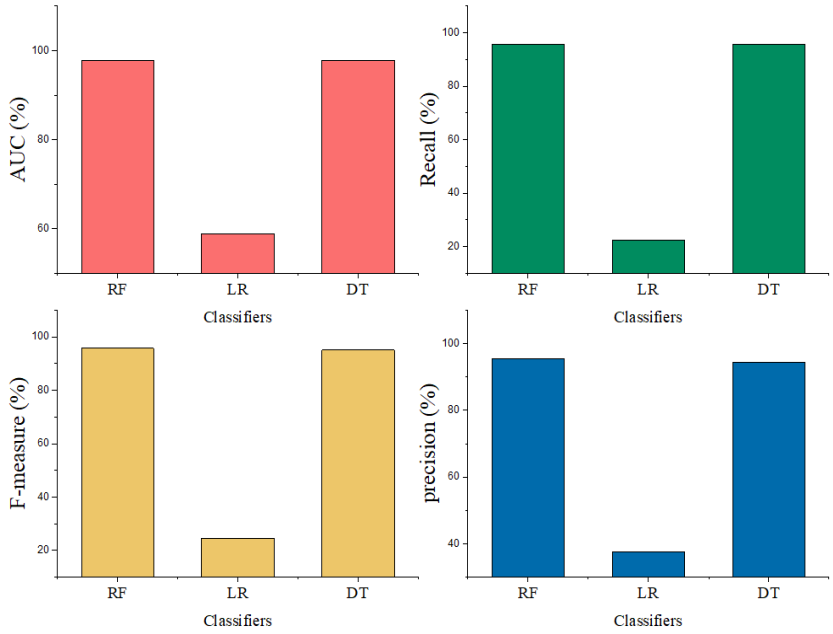

Fig 4: Performance of different classifiers

The sampling class is highly skewed; allocation is a difficulty involved. Numerous everyday machine learning and statistical classification techniques are prone to repetition bias, making learning how to discriminate between majority and minority classes difficult. To tackle the imbalanced class distributions in deep neural concepts, Fernando and Tsokos [32] presented a class rebalancing technique in the basis of a dynamically balanced class with weights assigned according to the class occurrence and the anticipated likelihood of regression coefficients class. Considering the dynamic weighted scheme's capacity to consciously its values in response to option to ensure, the system may be adjusted for situations of varying complexity, resulting in gradients updates powered by complex class samples.

The introduction of Gradient boost (GB) is for the ML classifiers and the regression models. Besides, GB is also the one type of Ml, which is helped to boost the classification parameters [35]. When the GB is applied in any classifier model, it has tuned the classification parameter to the desired level. So, the optimized output will be gained. In addition, XGboost (XGB) also combined with the GB to offer the portable sufficient library function that has helped to execute the GB on all platforms for different applications and its purposes [36]. For the smoothening process, the cross-entropy (CE) models were used along with the neural approaches. Based on the neural approaches parameters, the functioning and process of the CE will differ [37].

Multi-label categorization is inherently unbalanced in terms of data. Despite many studies, the class imbalance problem has continued to provide a hurdle for inter-categorization. Consider the sixteen-label categorization job. The subset of the identifiers contains several potential possibilities. As a result, it is impossible to acquire a balanced database for any label combination. Numerous researches on inter-categorization either attempt to balance the dataset by rescaling or disregarding the imbalance. However, because the under- and over-sampling approaches were not developed for multi-label categorization, it is not easy to adapt them to this environment. One frequently used heuristic is the use of the inverse subclass frequency per weighting category. Moreover, the Dynamically Weighted (DW) model can

balance the training loss when applied to ML approaches [33].

Table 1: Merits and demerits of discussed literatures**.**

Short summary			
Authors	**techniques**	**Advantages**	**Disadvantages**
Yao et al., [30]	DCNM	The skin-lesion datasets is used as the imbalanced data and has gained the desired specification outcome	But, the skin data is too complex. So more resources were required for the execution process
Wan et al., [31]	upgraded VGG19	The attained exactness score for classification is 97.8%	It is suitable only for large datasets
Yan and Wen [35]	GB	Based on the dataset complexity, the gradients were upgraded	Libraries were limited
Nguyen et al., [36]	XGB	It has afforded the support for GB mechanism by offering the different libraries	When it executed independent basis, less performance score was gained
Yilmaz et al.,[33]	DW	It can reduce the effect of CE loss. Hence, the error get reduced	Design complexity
Dong et al., [37]	CE	To measure the error in each data point, the CE approaches were utilized	But in complex data, error prediction is very less

In addition, the modified Focal loss has adopted both the cross and dice entropy based on losses to switch the class imbalance data. However, dual key issues were associated with the modified Focal loss techniques in practice [34]. To make the data balanced, cross-entropy (CE) is very useful, validating the possible error occurrences in each data point. But, if the data is unstructured, the error calculation rate becomes less [35]. The merits and limitations of the discussed literature are tabulated in Table.1.

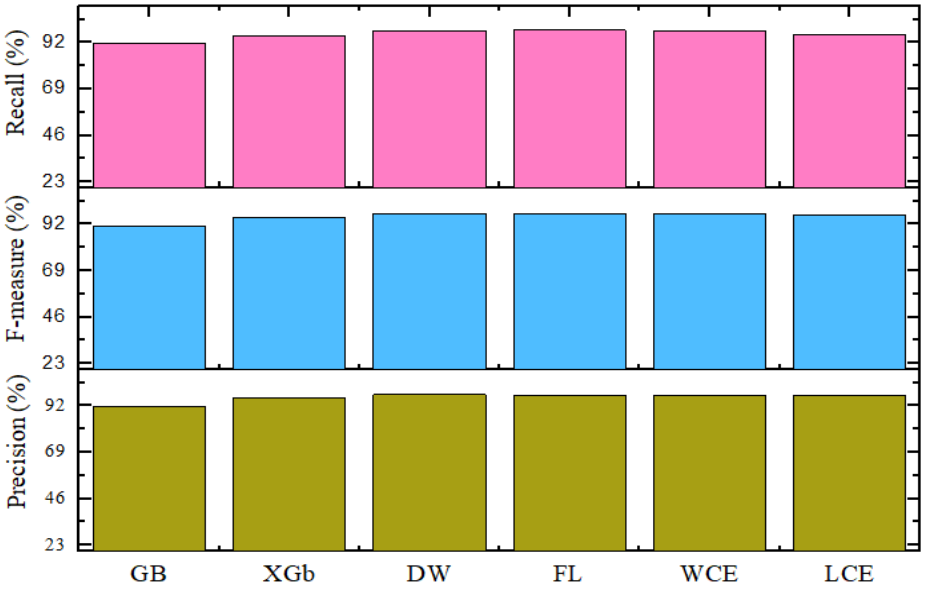

Fig 5: Performance of boosting techniques

Several DL concepts were discussed for the imbalanced data classification. Hence, the robustness and efficiency of each boosting model are verified by calculating the key parameters demonstrated in Figure 5. In addition, the few merits and limitations of the discussed related research works are tabulated in Table 1.

2.3 Optimization Models in imbalanced data classification

It is challenging to design an appropriate classifier for elevated unbalanced data, which significantly impairs the efficiency of classifiers [44]. Though numerous ways have been developed to cope with flawed statistics, such as oversampling, cost-sensitive, and ensembles learning models, they are hampered by comprehensive data with distortion and redundancy. So, to overcome restrictions, an adaptable subspace optimal ensemble approach (ASOEM) for high-dimensional unbalanced data classification [38]. To produce various resilient and discriminative embedding, a new adaptive-subspace optimization (ASO) approach and rotational-subspace-optimization (RSO) is devised. The optimized subspace is again resampled to create class-balanced information for each classifier. To demonstrate its effectiveness of the developed ASOEM,different experimental trials have been conducted. Hassib et al., [39] have introduced the Whale optimization (WO) for the imbalanced data specification. One of the fancy creatures in the sea world is the whale. They are recognized as the largest animals on the planet. They were discussed for the imbalanced data classification. Hence, the robustness and efficiency of each boosting model are verified by calculating the critical parameters demonstrated in Figure 5. In addition, the few merits and limitations of the discussed related research works are tabulated in Table1.

There are seven different primary species of this giant animal: Minke, killer, Sei, right, humpback, blue, and finback. Whales are typically viewed as predators. In addition, it never sleeps even though they breathe from the sea level. In reality, half portion of the brain only sleeps. Another intriguing issue is the whale's social activity; they can live in groups or alone. Often, they are generally spotted in groups. Here, this social behavior and hunting fitness has been utilized for the imbalanced data classification process. The WO has yielded the highest AUC of 99% by the validation. But it has recorded the average categorization exactness score as 81%. In addition, the imbalanced data has produced more complexity for the classification process. So, the data must be balanced before initiating the data mining or other prediction process. So, Shaw et al., [41] have introduced a particle Swarm algorithm (PSA) for the majority class specification purpose. Hence, the performance of the SA has been tested with 15 real-time imbalanced databases. The swarm model has gained a poor outcome in many cases during the results. So, the Ring theory with PSA (RTSA) [41] has been executed, and the parameters were noted. Also, the Ring theory on evaluation-learning (RTEL) [40] also checked with the real-time databases.

Fig 6: Summary of optimization models

Simulated-annealing (SA) is a well-known meta-heuristic for local search frequently used to solve intermittent and lesser degree continuous optimization issues. Moreover, the SA has the key advantage, enabling the local optima to use permitting hill-climbing movements to attain the global optimum. This optimal solution is utilized for the imbalanced data categorization issues [42]. Hence, the recorded G-mean score is 96.63. The summary of the optimization models is described in Figure 6, and the performance is exposed in Table 2.

Table 2: Performance of optimization models

S.no	Algorithms	Accuracy (%)	F-score (%)
1	WO	81	-
2	RTSA	-	97.5
3	PSA	-	100
4	RTEL	-	94.2
5	SA with DT	98.3	100
6	SA with SVM	96/5	71.46
7	SA with KNN	97.12	72.59
8	RO	90	90

Customer loyalty is critical for forecasting, and telecommunications industries are inextricably linked to financial institutions. As a result, several industries took a variety of activities to establish a strong connection with their customers and minimize user defections [43]. To achieve the highest possible customer loyalty, the essential learning parameters are the shifting consumer hierarchy and the factors that contribute to churning. Besides, Pustokhina et al., [43] have simulated the rain optimization (RO) for the customer churn prediction imbalanced data.

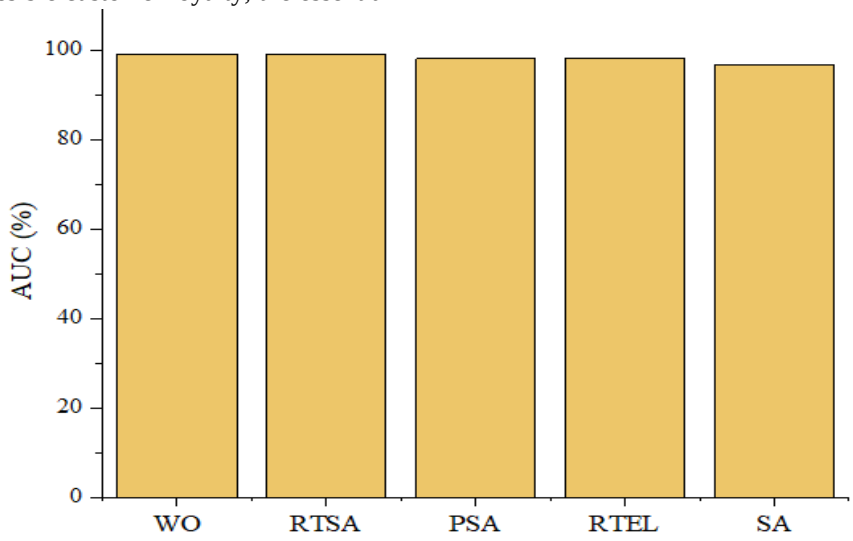

Fig 7: Validation of AUC

RO is a novel heuristic approach influenced by how raindrops gravitate toward minimal spots upon impact with the earth. In addition, if this method's parameters are tweaked properly, it can identify both global and local extremes. The AUC of the discussed literature is exposed in Figure 7.

3. Performance analysis

Several methods were discussed for the classification of the imbalanced data. Hence, different types of data were used that included several fields like medical agriculture, business, e-learning, etc. The overall analysis of the discussed models is detailed in Table.3.

Table 3: Performance analysis of different imbalanced data classification models

Overall review analysis							
Authors	**Method**	**Classification models**	**Accuracy**	**precision**	**F-measure**	**sensitivity**	AUC
Vuttipittayamongkol et al., [25]			99%		90	60	-
Sambasivam and Opiyo [27]	Convolution nets (CN)	Oversampling (OS)	93	92	92	91	-
Rupapara et al., [28]	Regression vector	RF	92	94	83	78	-
	Regression vector	KNN	89	86	78	74	-
	Regression vector	DT	91	84	85	85	-
	Regression vector	LR	94	91	89	87	-
	Regression vector	RV-VC	93	91	88	85	-
Maulidevi and Surendro [29]	SMOTE	C4.5	71	58	59	60	72.62
	SMOTE	Naïve bayes (NB)	76	67	63.80	60	81.69
	SMOTE	SVM	77.2	73.4	62.5	54.5	71.94
Yao et al., [30]	Multi weighted (MW)	Weighted models	86		97	78	97
	cumulative learning CL	-	87		98	81	98
Wan et al., [31]	improved VGG19	-	97.7	97	97	97	-
Fernando and Tsokos [32]	Gradient Bossting (GB)	Boosting parameters	-	91.5	91	91.3	95.6
	XGboost	Boosting parameters	-	95.9	95.5	95.3	97.6
	Dynamically Weighted (DW) loss	Weighted models	-	97.5	97.4	98	98.9
	Focal loss (FL)	Weighted models	-	96.96	97.4	98.05	98.8
	Weighted (WCE)	Weighted models	-	97.38	97.4	97.44	98.6
	Loss CE (LCE)	Entropies	-	97.15	96.5	96	97.9
Pustokhina et al., [43]	RO	-	90	91	90	88	-

Each technique has specific characteristics based on the classification rate determined. In addition, the imbalanced data is more complex in the data mining field; hence an efficient neural model was required to categorize the imbalanced data. In addition, the effectiveness of each approach is validated by estimating the key parameters such as accuracy, F-measure, precision, recall, and AUC.

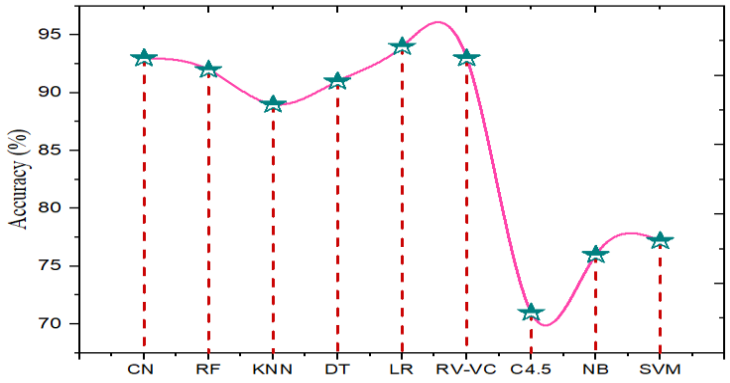

Fig 8: Accuracy comparison

The parameter accuracy has been estimated to measure the classification exactness score of each dataset. The accuracy has been calculated based on the value divided by the total samples. The accuracy validation of a few models is described in Figure 8. Hence, the accuracy formulation is equated in Eqn.(1), and the precision formulation is described in Eqn. (2).

$$Accuracy = \frac{True\_positives + True\_negative}{Total \quad Samples} \tag{1}$$

$$Precision = \frac{True\_positives}{False\_positives + True\_positives} \tag{2}$$

To measure the positive values in prediction scenario, the metrics precision was calculated.

$$recall = \frac{True \quad Positives}{false \quad negative + \quad true \quad positives} \tag{3}$$

The other name of recall is sensitivity; validating the sensitivity measures the robustness score based on different datasets. If the model has the highest recall validation, it has the finest classification exactness score; the formulation is described in Eqn. (3).

$$F - score = 2 \times \frac{recall \times precision}{\mathrm{Re}\, call + precision} \tag{4}$$

To find the mean value of the metrics recall and precision, the F-measure parameters were measured, which is formulated in Eqn.(4).

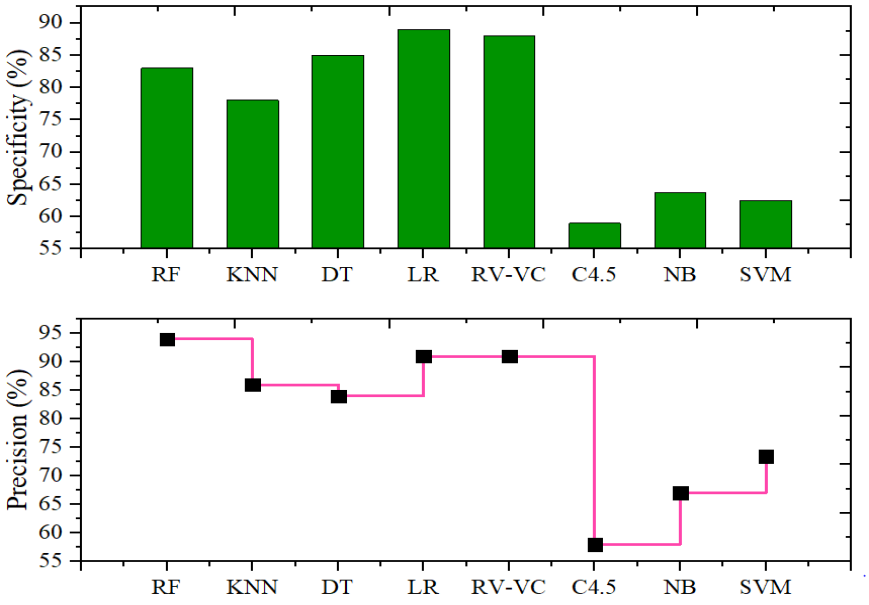

Fig 9: Precision and specificity validation

Here, the RF approach has earned the imbalanced data specification exactness score of 92%, precision 94%, F-measure 83%, and recall 78%. The KNN model has gained 89% accuracy for imbalanced data categorization, 86% precision, F-score 78%, and remembers 74%. The approach DT Has yielded an imbalanced data categorization score of 91%, precision 84%, F-measure 85%, and sensitivity 85%. [46] Moreover, the LR scheme has achieved an imbalanced data classification precision score of 91%, an accuracy of 94%f-score of 89%, and a sensitivity of 87%. The RV0VC model has recorded the imbalanced data classification exactness score as 93%, precision 91%, sensitivity 85%, and f-measure 88%. The C4,5 scheme has yielded the accuracy for imbalanced data classification is 71%, precision of 58%, F-score of 59%, and recall of 60%. The model NB has gained the exactness score for imbalanced data specification is 76%, 67% precision, 63.80 F-score, and 60% sensitivity. Finally, the model support-vector-model (SWM) has gained 77.25% accuracy for imbalanced data classification, 73.4% precision, 62.5% f-score, and 54.5% recall. Overall, the knowledge-based system [25] has gained the finest accuracy as 99%. Precision and specificity validation are shown in Figure 9.

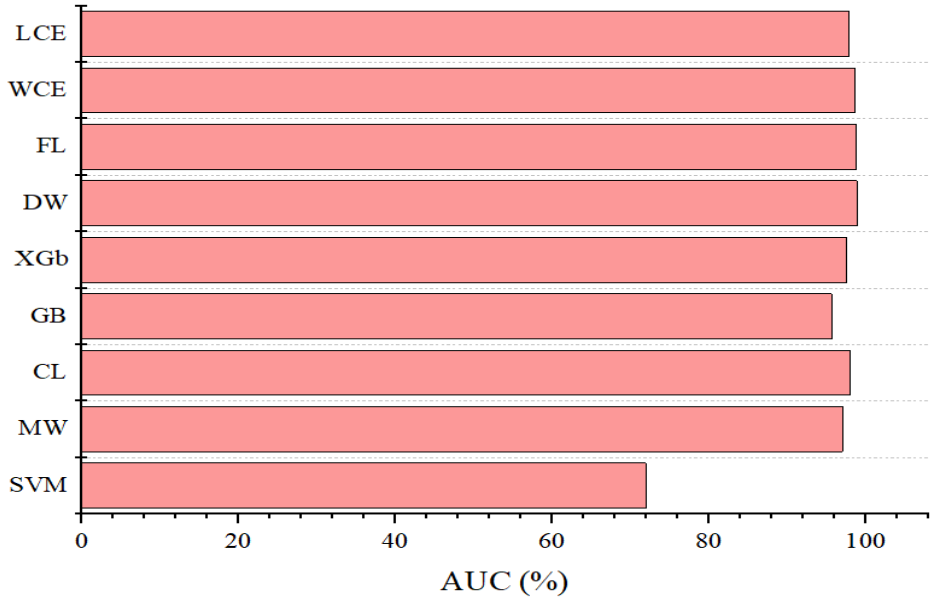

Fig 10: AUC comparison validation

To measure the positive values in the prediction results, the metrics Area-Under-Curve (AUC) has been counted. If the model has gained the finest AUC rate, it is good in the prediction function. Hence, the validation of AUC is explained in Figure 10.

4. Discussion

The main motive of this review is to analyze the imbalanced data classification techniques with strengths and weaknesses. Hence, some optimization models have gained the 100% F-measure after reviewing several methods. Moreover, the models are SA and PSA. However, the two algorithms haven't attained the 100% F-score for all datasets based on the data complexity and unstructured rate; the F-score might differ. Also, those approaches have required more time to complete all iterations till the desired solution is met.

5. Conclusion

This review article aims to find the future direction for imbalanced data classification by reviewing the different existing classification algorithms. Also, the efficiency of the discussed literature was identified using some key parameters like AUC, precision, F-measure recall, and accuracy. Hence, by comparing these metrics, the model SA has gained the 100% f-measure, which shows it fits the imbalanced data classification applications. Several ML and DL approaches have achieved the best outcome in classifying the imbalanced data. In addition, different classifiers are also utilized in different ways by the combination of oversampling techniques and optimizations. However, many algorithms were lacking in resources usage. So, in the future, designing hybrid DL models with hybrid optimization models will provide the

multi-functions employed for multi- imbalanced data classification with finest exactness score.

References

[1] X. Yin, Q. Liu, Y. Pan, X. Huang, J. Wu, and X. Wang, 2021. Strength of stacking technique of ensemble learning in rockburst prediction with imbalanced data: Comparison of eight single and ensemble models. *Natural Resources Research*, *30*, pp.1795-1815.

[2] A. Dogan, and D. Birant, 2021. Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*, p.114060.

[3] H. Thakkar, V. Shah, H. Yagnik, and M. Shah, 2021. Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis. *Clinical eHealth*, *4*, pp.12-23.

[4] Y. Pan, and L. Zhang, 2021. A BIM-data mining integrated digital twin framework for advanced project management. *Automation in Construction*, *124*, p.103564.

[5] P. Espadinha-Cruz, R. Godina, and E.M. Rodrigues, 2021. A review of data mining applications in semiconductor manufacturing. *Processes*, *9*(2), p.305.

[6] J. Jedrzejowicz, and P. Jedrzejowicz, 2021. GEP-based classifier for mining imbalanced data. *Expert Systems with Applications*, *164*, p.114058.

[7] P. Liu, W. Qingqing, and W. Liu, 2021. Enterprise human resource management platform based on FPGA and data mining. *Microprocessors and Microsystems*, *80*, p.103330.

[8] K.G. Al-Hashedi, and P. Magalingam, 2021. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, p.100402.

[9] Z. Sanad, and A. Al-Sartawi, 2021, March. Financial statements fraud and data mining: a review. In *European, Asian, Middle Eastern, North African Conference on Management & Information Systems* (pp. 407-414). Cham: Springer International Publishing.

[10] L. Shabtay, P. Fournier-Viger, R. Yaari, and I. Dattner, 2021. A guided FP-Growth algorithm for mining multitude-targeted item-sets and class association rules in imbalanced data. *Information Sciences*, *553*, pp.353-375.

[11] E. Aminian, R.P. Ribeiro, and J. Gama, 2021. Chebyshev approaches for imbalanced data streams regression models. *Data Mining and Knowledge Discovery*, *35*, pp.2389-2466.

[12] L. Korycki, and B. Krawczyk, 2021, May. Low-dimensional representation learning from imbalanced data streams. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 629-641). Cham: Springer International Publishing.

[13] J. Grzyb, J. Klikowski, and M. Woźniak, 2021. Hellinger distance weighted ensemble for imbalanced data stream classification. *Journal of Computational Science*, *51*, p.101314.

[14] N. Lu and T. Yin, 2021. Transferable common feature space mining for fault diagnosis with imbalanced data. *Mechanical systems and signal processing*, *156*, p.107645.

[15] D. Sisodia, and D.S. Sisodia, 2022. Data sampling strategies for click fraud detection using imbalanced user click data of online advertising: an empirical review. *IETE Technical Review*, *39*(4), pp.789-798.

[16] B. Mirzaei, B. Nikpour, and H. Nezamabadi-pour, 2021. CDBH: A clustering and density-based hybrid approach for imbalanced data classification. *Expert Systems with Applications*, *164*, p.114035.

[17] S.X. Chen, X.K. Wang, H.Y. Zhang, and J.Q. Wang, 2021. Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine. *Expert Systems with Applications*, *173*, p.114756.

[18] S. Zhu, 2021. Analysis of the severity of vehicle-bicycle crashes with data mining techniques. *Journal of safety research*, *76*, pp.218-227.

[19] K. Yang, Z. Yu, C.P. Chen, W. Cao, J. You, and H.S. Wong, 2021. Incremental weighted ensemble broad learning system for imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *34*(12), pp.5809-5824.

[20] G.A. Pradipta, R. Wardoyo, A. Musdholifah, and I.N.H. Sanjaya, 2021. Radius-SMOTE: a new oversampling technique of minority samples based on radius distance for learning from imbalanced data. *IEEE Access*, *9*, pp.74763-74777.

[21] W. Wang, and D. Sun, 2021. The improved AdaBoost algorithms for imbalanced data classification. *Information Sciences*, *563*, pp.358-374.

[22] C. Hou, J. Wu, B. Cao, and J. Fan, 2021. A deep-learning prediction model for imbalanced time series data forecasting. *Big Data Mining and Analytics*, *4*(4), pp.266-278.

[23] R.M. Pereira, Y.M. Costa, and C.N. Silla Jr, 2021. Toward hierarchical classification of imbalanced data using random resampling algorithms. *Information Sciences*, *578*, pp.344-363.

[24] X. Wang, J. Xu, T. Zeng, and L. Jing, 2021. Local distribution-based adaptive minority oversampling for imbalanced data classification. *Neurocomputing*, *422*, pp.200-213.

[25] P. Vuttipittayamongkol, E. Elyan, and A. Petrovski, 2021. On the class overlap problem in imbalanced data classification. *Knowledge-based systems*, *212*, p.106631.

[26] L.M. Dang, S. Kyeong, Y. Li, H. Wang, T.N. Nguyen, and H. Moon, 2021. Deep learning-based sewer defect classification for highly imbalanced dataset. *Computers & Industrial Engineering*, *161*, p.107630.

[27] G.A.O.G.D. Sambasivam, and G.D. Opiyo, 2021. A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks. *Egyptian informatics journal*, *22*(1), pp.27-34.

[28] V. Rupapara, F. Rustam, H.F. Shahzad, A. Mehmood, I. Ashraf, and G.S. Choi, 2021. Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. *IEEE Access*, *9*, pp.78621-78634.

[29] N.U. Maulidevi, and K. Surendro, 2022. SMOTE-LOF for noise identification in imbalanced data classification. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), pp.3413-3423.

[30] P. Yao, S. Shen, M. Xu, P. Liu, F. Zhang, J. Xing, P. Shao, B. Kaffenberger, and R.X. Xu, 2021. Single model deep learning on imbalanced small datasets for skin lesion classification. *IEEE transactions on medical imaging*, *41*(5), pp.1242-1254.

[31] X. Wan, X. Zhang, and L. Liu, 2021. An improved VGG19 transfer learning strip steel surface defect recognition deep neural network based on few samples and imbalanced datasets. *Applied Sciences*, *11*(6), p.2606.

[32] K.R.M. Fernando, and C.P. Tsokos, 2021. Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(7), pp.2940-2951.

[33] S.F. Yilmaz, E.B. Kaynak, A. Koç, H. Dibeklioğlu, and S.S. Kozat, 2021. Multi-label sentiment analysis on 100 languages with dynamic weighting for label imbalance. *IEEE Transactions on Neural Networks and Learning Systems*.

[34] Y. Kim, Y. Lee, and M. Jeon, 2021. Imbalanced image classification with complement cross entropy. *Pattern Recognition Letters*, *151*, pp.33-40.

[35] Z. Yan, and H. Wen, 2021. Electricity theft detection base on extreme gradient boosting in AMI. *IEEE Transactions on Instrumentation and Measurement*, *70*, pp.1-9.

[36] H.T.T. Nguyen, L.H. Chen, V.S. Saravanarajan, and H.Q. Pham, 2021, May. Using XG Boost and Random Forest Classifier Algorithms to Predict Student Behavior. In *2021 Emerging Trends in Industry 4.0 (ETI 4.0)* (pp. 1-5). IEEE.

[37] Y. Dong, X. Shen, Z. Jiang, and H. Wang, 2021. Recognition of imbalanced underwater acoustic datasets with exponentially weighted cross-entropy loss. *Applied Acoustics*, *174*, p.107740.

[38] Y. Xu, Z. Yu, C.P. Chen, and Z. Liu, 2021. Adaptive subspace optimization ensemble method for high-dimensional imbalanced data classification. *IEEE Transactions on Neural Networks and Learning Systems*.

[39] E.M. Hassib, A.I. El-Desouky, L.M. Labib, and E.S.M. El-Kenawy, 2020. WOA+ BRNN: An imbalanced big data classification framework using Whale optimization and deep neural network. *soft computing*, *24*, pp.5573-5592.

[40] Z. Li, Q. Zhang, and Y. He, 2022. Modified group theory-based optimization algorithms for numerical optimization. *Applied Intelligence*, *52*(10), pp.11300-11323.

[41] S.S. Shaw, S. Ahmed, S. Malakar, L. Garcia-Hernandez, A. Abraham, and R. Sarkar, 2021. Hybridization of ring theory-based evolutionary algorithm and particle swarm optimization to solve class imbalance problem. *Complex & Intelligent Systems*, *7*(4), pp.2069-2091.

[42] A.S. Desuky, and S. Hussain, 2021. An improved hybrid approach for handling class imbalance problem. *Arabian Journal for Science and Engineering*, *46*, pp.3853-3864.

[43] I.V. Pustokhina, D.A. Pustokhin, P.T. Nguyen, M. Elhoseny, and K. Shankar, 2021. Multi-objective rain optimization algorithm with WELM model for customer churn prediction in telecommunication sector. *Complex & Intelligent Systems*, pp.1-13.

[44] J.F. Cui, H. Xia, R. Zhang, B.X. Hu, and X.G. Cheng, 2021. Optimization scheme for intrusion detection scheme GBDT in edge computing center. *Computer Communications*, *168*, pp.136-145.

[45] I.V. Pustokhina, D.A. Pustokhin, R.H. Aswathy, T. Jayasankar, C. Jeyalakshmi, V.G. Díaz,. and K. Shankar, 2021. Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms. *Information Processing & Management*, *58*(6), p.102706.

[46] J. Yao, Y. Zheng, and H. Jiang, 2021. An ensemble model for fake online review detection based on data resampling, feature pruning, and parameter optimization. *Ieee Access*, *9*, pp.16914-16927.

[47] Ravi, C., Yasmeen, Y., Masthan, K. ., Tulasi, R. ., Sriveni, D. ., & Shajahan, P. . (2023). A Novel Machine Learning Framework for Tracing Covid Contact Details by Using Time Series Locational data & Prediction Techniques. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2s), 204–211. https://doi.org/10.17762/ijritcc.v11i2s.6046

[48] Rossi, G., Nowak, K., Nielsen, M., García, A., & Silva, J. Enhancing Collaborative Learning in Engineering Education with Machine Learning. Kuwait Journal of Machine Learning, 1(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/119

[49] Veeraiah, D., Mohanty, R., Kundu, S., Dhabliya, D., Tiwari, M., Jamal, S. S., & Halifa, A. (2022). Detection of malicious cloud bandwidth consumption in cloud computing using machine learning techniques. Computational Intelligence and Neuroscience, 2022 doi:10.1155/2022/4003403