Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Imbalance Dataset Handling for Classification using Machine Learning Algorithm

Dr. Sanjay Kumar N V[1], Dr. Nanditha Krishna[2], Dr. Kavita Avinash Patil[3], Salna Joy[4], R Baby Chithra[4], Dr. Raghavendra Patil G E[5]

[1]Kalpataru Institute of Technology, Tiptur, Karnataka, India
[2]Center for Medical Electronics and Computational Intelligence, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India
[3]M. S. Ramaiah Institute of Technology, Bengaluru, Karnataka, India
[4]New Horizon College of Engineering, Bengaluru, Karnataka, India
[5]Jain Deemed-to be University, Bengaluru, Karnataka, India

Machine learning models are severely impacted by unbalanced data. Models may become biased in favor of the majority class when one class has a significantly larger sample count than the other, which would hurt the minority class's performance. Overfitting is made more likely by imbalanced data since the model may become accustomed to memorizing most class samples rather than identifying underlying patterns. In order to tackle these problems in the classification space, this work investigates a number of approaches, such as cost-sensitive learning, SMOTE, under- and oversampling, and ensemble deep learning techniques. They assess how well these techniques work with various datasets and offer details on their advantages and disadvantages. A taxonomy of approaches, including as resampling, algorithmic and other approaches, for imbalanced binary and multi-class classification issues is presented in the paper.
Keywords: Machine learning, Imbalanced data, Smote, Classification space.

1. Introduction

The main issue in the current situation is unbalanced data. Managing unbalanced data in a computing process is a challenging issue. If there are more instances of one class in a data collection than the other, that class is said to be a majority class. Stated otherwise, a class is considered minority if there are fewer instances of it in a database than there are of another class in the same database. We refer to these types of data sets as imbalanced data sets. One of the key functions of pattern recognition is classification. Several methods for classification

learning, include support vector machines, decision trees, backpropagation neural networks, Bayesian networks, nearest neighbors, and the recently have been developed for image face detection. The datasets used for classification tasks in many real-world contexts are generally imbalanced, indicating that one class has a much higher number of occurrences than the others. Machine learning models have difficulties when dealing with imbalanced datasets because they often exhibit bias towards the majority class, which results in subpar performance on minority classes. This problem is widespread in many different fields, such as anomaly detection, fraud detection, and medical diagnostics. To effectively address the imbalance issue and make sure the model learns from both majority and minority classes, specific strategies are needed. We examine the methods and techniques frequently used to address imbalanced classification challenges with machine learning algorithms in this paper, chapter, or report. We investigate strategies like cost-sensitive learning, ensemble approaches, resampling, and algorithmic changes intended to lessen the effects of class imbalance.

2. LITERATURE SURVEY

Nitesh V. Chawla et.al [1] introduced SMOTE, a seminal method for over- sampling the minority class by generating synthetic examples along line segments joining any/all minority class nearest neighbors. Edwardo A. Garcia et.al [2] provided a comprehensive overview of techniques for handling imbalanced datasets, including resampling methods, cost-sensitive learning, and algorithm- specific approaches. A. Fernández et.al [3] investigated the effectiveness of Random Forests for classification on imbalanced datasets and proposes techniques for improving its performance in such scenarios. Pablo M. Olmos et.al [4] proposed a hybrid approach that combines SMOTE for over-sampling and ensemble classifiers to address imbalanced datasets effectively. Longbing Cao et.al [5] introduces a cost-sensitive boosting method tailored for imbalanced datasets, which adjusts the misclassification costs during the boosting process to alleviate class imbalance. Michele Filannino et.al [6] discusses techniques for handling both data sparsity and class imbalance in the context of predicting the popularity of online news articles.

3. IMBALANCE DATA AND HANDLING METHODS

Imbalanced data is a dataset where the classes are not equally represented. This can be a problem in machine learning because traditional models tend to perform poorly on the minority class. In other words, Imbalanced data refers to a dataset where the classes are not equally represented. For example, in a binary classification problem with two classes (class A and class B), if the number of samples belonging to class A is significantly larger than the number of samples belonging to class B, the dataset is imbalanced.

1. Oversampling: A Simple Trick to Balance Your Data

In the world of data analysis, it's crucial to have a balanced set of data to get accurate results. "Oversampling" is a handy technique to achieve this balance. It involves making copies of the smaller group of data until it's as large as the bigger group. This way, the analysis is fair and gives us better results.

2. Under-sampling: A Strategy to Balance Your Data by Reducing Excess

In data analysis, sometimes we have too much information from one group, which can skew the results. A technique called "under-sampling" can help fix this. It means we remove some examples from the larger group until it matches the size of the smaller group. This ensures a balanced view, helping to create a more reliable analysis.

3. Hybrid Sampling: Mixing Two Methods for Better Data When we work with data, it's important to have a balanced mix to get the best results. "Hybrid sampling" helps us do just that. It's a method where we add more examples to the smaller group and take away some from the bigger group. This way, we have a fair and even set of data to work with.

4. SMOTE: Creating New Examples to Balance Data

While analyzing the data sometimes we need to add more examples to the smaller group to make the data balanced. "SMOTE", which stands for Synthetic Minority Oversampling Technique, is a method that helps us do this. It picks two similar examples from the smaller group and makes a new example that is a mix of the two.

5. ADASYN: Adjusting the Data Balance Smartly

When we are dealing with data, sometimes the groups are not evenly matched, with one group having much more data than the other. "ADASYN", Sampling, is a smart tool that helps to even things out. It creates new examples in the smaller group, and the number of new examples it makes depends on how imbalanced the data is to start with.

6. Tomek Links: A Method to Fine-Tune Your Data Sometimes while analyzing the data necessary to remove some data points to get a clearer picture. "Tomek Links" is a technique that helps us do this. It finds pairs of data where one is from the larger group and the other is from the smaller group, and they are very similar. Then, it removes the data point from the larger group to make the data more balanced.

7. ENN Rule: Cleaning Up Data with the Help of Neighbors When analyzing data, it's crucial to have a clean and balanced dataset to get accurate results. The "Edited Nearest Neighbor" or ENN rule helps us achieve this. It works by spotting and removing data points from the larger group that are similar to points in the smaller group, helping to clear up any confusion and make the data more reliable.

8. Condensed Nearest Neighbor Rule: Building a Focused Dataset

In data analysis, it's often beneficial to create a focused dataset that zeroes in on the most relevant examples. The "Condensed Nearest Neighbor" rule helps us do this by forming a new dataset that includes only the most relevant examples from the larger group, along with all the examples from the smaller group. This way, we can concentrate on the data points that matter the most.

9. Near Miss Method: A Focused Approach to Balancing Imbalanced Datasets

Balancing imbalanced datasets is a crucial step in data analysis, especially when there is asignificant disparity between the classes. The near-miss method is a technique designed to streamline this process. It works by identifying and removing data points from the majority class that are closely similar to those in the minority class, thus helping to create a more

balanced and focused dataset.

10. One-Sided Selection: Balance the Imbalanced Data

One-sided selection is a technique for oversampling the minority class in an imbalanced dataset. It involves selecting a subset of the majority class that is most similar to the minority class. Then the rest of the majority class examples are removed from the dataset. To do this, we first identify the majority class examples that are nearest to minority class examples. We can use a variety of methods to do this, such as k-nearest neighbors or k-means clustering.

4. METHODOLOGY

As previously stated, data level solutions encompass a wide range of resampling techniques, including directed oversampling, directed under-sampling, random oversampling with replacement, random under-sampling, oversampling with informed generation of new samples, and combinations of the afore mentioned methods.

(i)Under sampling

By randomly removing instances from the majority class, random under-sampling is a non-heuristic technique that seeks to balance the distribution of classes. The idea behind it is to attempt dataset balancing as a means of mitigating the peculiarities of the machine learning algorithm. Random under sampling's main flaw is that it may miss out on information that could be crucial for the induction process but could still be valuable. This method also has a flaw in that machine learning is meant to help the classifier determine the probability distribution of the intended audience. We attempt to use a sample distribution to estimate the population distribution because that distribution is unknown.

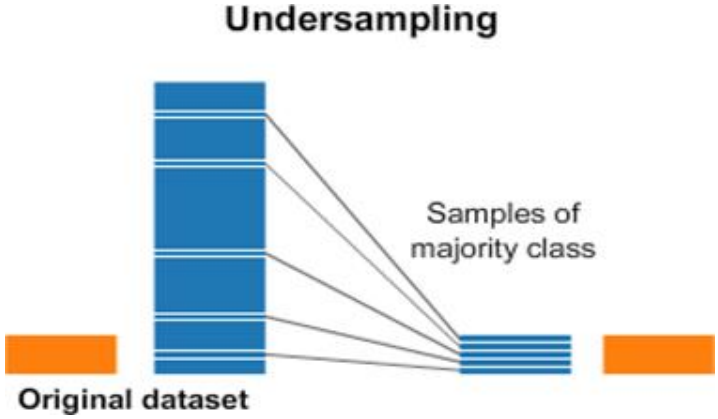

Fig. 1 Undersampling process

According to statistics, the sample distribution can be used to estimate the population distribution from which the sample was collected as long as it was chosen at random. Therefore, we can learn to estimate the target distribution by studying the sample distribution. could still be valuable. This method also has a flaw in that machine learning is meant to help the classifier determine the probability distribution of the intended audience. We attempt to

use a sample distribution to estimate the population distribution because that distribution is unknown [7]. According to statistics, the sample distribution can be used to estimate the population distribution from which the sample was collected as long as it was chosen at random. Therefore, we can learn to estimate the target distribution by studying the sample distribution. In this approach, we reduce the number of samples from the majority class to match the number of samples in the minority class [8].

(ii)Over sampling

The goal of random over-sampling, a non-heuristic technique, is to replicate minority class examples at random in order to achieve class distribution parity. Many writers, concur that because random oversampling creates exact duplicates of the minority class samples, it may raise the risk of overfitting. In this sense, a symbolic classifier, for example, could create rules that seem accurate but only cover one case that is repeated [9]. In addition, if the data set is already somewhat large but unbalanced, oversampling may result in an extra computational bur In order to oversample the minority class, SMOTE creates artificial minority examples. The fundamental idea behind it is to create new examples of minority classes by creating interpolations between examples of minority classes that are close together. In this approach, we synthesize new examples from the minority class. There are several methods available to oversample a dataset used in a typical classification problem [10]. But the most common data augmentation technique is known as Synthetic Minority Oversampling Technique or SMOTE for short.

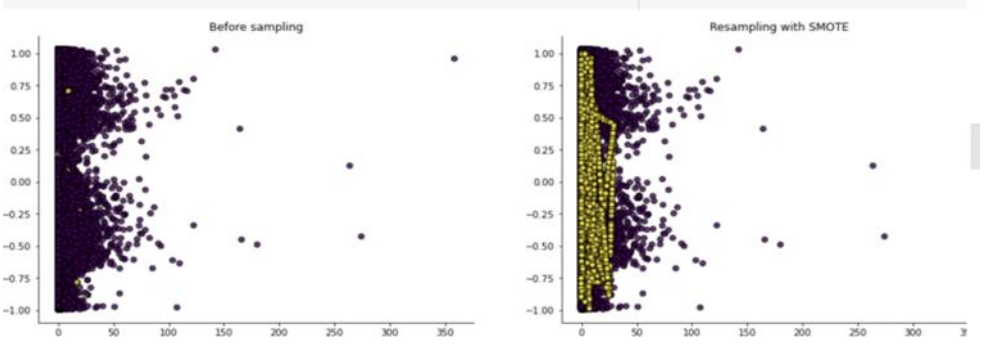

Fig. 2 Oversampling process

(iii)SMOTE

As the name suggests, SMOTE creates “synthetic” examples rather than over-sampling with replacement. Specifically, SMOTE works the following way. It starts by randomly selecting a minority class example and finding its k nearest minority class neighbors at random. Then a synthetic example is created at a randomly selected point in the line that connects two examples in feature space.

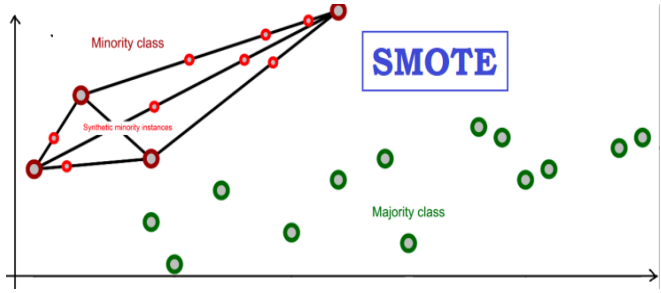

Fig 3. SMOTE Analysis

5. HYBRID APPROACH HANDLING CLASS IMBALANCE PROBLEM

In addition to ensemble approaches, cost-sensitive techniques, and one-class learning, a new generation of classification algorithms has emerged recently to address datasets with class imbalance. To enhance the quality of categorization, the majority of them utilize multiple machines learning algorithms, frequently combining them with other learning algorithms to get superior [11]. The purpose of hybridization is to reduce the difficulties associated with sampling, selecting feature subsets, optimizing cost matrices. A number of published studies, such as, have shown how to combine cost-sensitive learning with sampling using the SMOTE algorithm to increase SVM performance. Additionally, a reported study from proposed a cost-sensitive neural network based on PSOs.

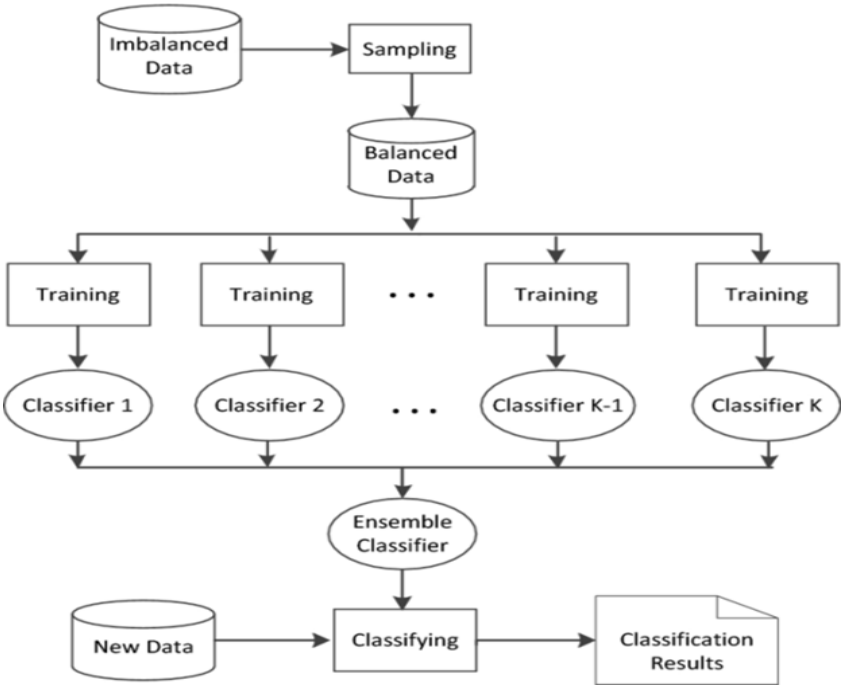

Fig. 4 Hybrid approach for imbalanced sampled data

Hybridization is an approach that exploits the strengths of individual components. When it comes to dealing with imbalanced classification data, some works proposed hybridization of sampling and cost-sensitive learning [12]. In other words, combining both data and algorithm level approaches. This idea of two-stage training that merges data- level solutions with algorithm-level solutions (i.e. classifier ensemble), resulting in robust and efficient learners is highly popular.

Table I: Hardware and software configuration of the test setup

Components	Remarks
Number of CPU Cores	16 CPUs Each of CPU is 64 Bit, 3.60GHz (Giga Hertz)
Main Memory Size	16dB (Giga Bites)
Hard Disk Size	SSD (Solid State Drive) of size 1TB (Tera Bytes)
Linux Kernel	Version 5.11.0
SBK	Version 0.97 [3]
Java Virtual Machine (JVM)	Version 17 [8]

ROC and AUC imbalanced data

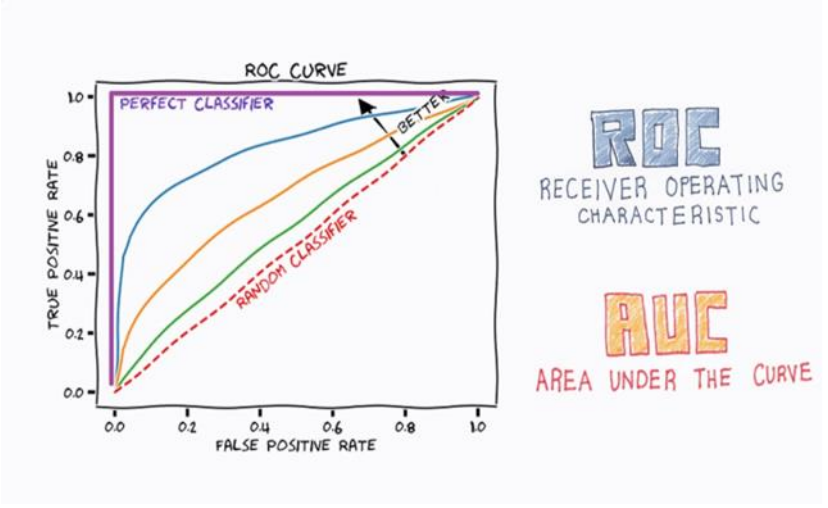

Fig. 5. ROC and AUC imbalanced data

To accommodate the minority class, the Receiver Operating Characteristic (ROC) curve is proposed as a measure over a range of trade-offs between the True Positive (TP) Rate and False Positive (FP) Rate [13]. Another important performance measure is Area Under the Curve

(AUC) is a commonly used performance metric for summarizing the ROC curve in a single score. Moreover, AUC is not biased towards the model's performance on either the majority or minority class, which makes this measure more appropriate when dealing with imbalanced data.

6. OTHER PROBLEMS RELATED WITH IMBALANCE

On the other hand, it has also been shown that traditional machine learning algorithms can provide good classifiers in certain domains, such as the Sick data set, even with training sets that are severely imbalanced. It can be shown from this that learning algorithms are not doing as well because of class imbalance alone. The distribution of data inside each class is also important (between-class versus within-class imbalance) therefore managing class imbalances is not the sole issue to deal with. Class imbalances may or may not impede classifier induction, and Prati et al.'s thorough investigation sought to determine whether these shortcomings could have alternative explanations. A multitude of studies examined the relationship between the class imbalance and other problems, including data duplication, overlapping classes, small disjunct and unusual cases complications. In certain instances, it was discovered that solving the small disjunct problem alone without taking into account the class imbalance issue was sufficient to improve performance. Less tuning is needed for this approach, which handles unusual case disjuncts similarly to the m-estimation Laplace smoothing [14]. Data duplication was also found to be detrimental in general, albeit significant levels of duplication are required to negatively impact classification for classifiers like Naive Bayes and Perceptrons with Margins [15]. The resampling approach that suggests is comprised of clustering the training data (individually) for every class, then randomly oversampling each cluster. The concept is to take into account both the within-class imbalance (the imbalance that exists within each class's subclusters) and the between- class imbalance (the imbalance that exists between the two classes). By addressing these two kinds of imbalances at the same time, you will oversample the dataset [16].

7. TRENDS IN CLASS IMBALANCE LEARNING AND CLASSIFICATION NOW AND IN THE FUTURE TRENDS

It is recognized in the research community that while there are more published works addressing the binary class imbalance problem, there are relatively few studies addressing the multiple class imbalance problem. This is most likely due to two factors: 1) the greater degree of complexity when dealing with multiple class imbalance problems, and 2) the prevalence of class unbalanced datasets with binary classes in most domain applications, such as anomaly detection to set the minority class apart from the prevailing class. The classifier is now burdened with learning the minority class from a very small training sample using a very high number of characteristics [17]. Even worse, very few classifiers are intended to manage a high volume of superfluous features. Due to redundancy problems—that is, characteristics that are partly or totally irrelevant to objective learning—there are numerous situations in which using all of the features that are available does not ensure that a superior classification output may be obtained.

Class overlapping issues can occasionally result in irrelevant features in the case of data with class imbalance. When data points fall within the zone where the majority class and minority class overlap, this situation arises. As not every feature may provide discriminant information that effectively distinguishes one class from an unfavorable class, the processing required to gather all the features is beginning to burden a classifier, particularly in terms of execution time and memory consumption. Effective feature selection is essential to support classification

jobs since processing large volumes of data with multivariate features requires minimizing computing costs [18]. The benefits of feature selection are numerous. They include the potential to reduce overfitting in predictions, speed up processes and make models more affordable in terms of storage and training times, and aid in the visualization of underlying data distribution in feature space for improved comprehension and resistance to the curve, One of the key obstacles in the classification field is still large data sets. Both the number and attribute/feature-wise expansion of data sets occurs quickly. Class imbalance is an expected issue with big data as the computing world moves toward this domain [19].

The features will expand into complex surfaces in multidimensional space when a large number of characteristics are required to describe a single data point. Knowledge extraction and data mining are made more difficult when there are complicated and non-linear relationships between multiple features. In [20][21][22], a number of papers on imbalanced large data are published, the majority of which concentrated on creating machine learning algorithms using the MapReduce architecture. It is anticipated that growing real-world demand for big data applications will demand for more advancements in machine learning algorithms to address the uneven handling of large amounts of data.

Future classification tasks will likely be shaped by the big data computing industry's rapid development, and since anomalous patterns are present in the majority of real-world problems, class imbalance problems are unavoidable. Motivated by the aforementioned considerations, it's also noteworthy to observe that new frameworks [23] and systematic mapping [24] will likely be developed as a possible study topic when machine learning and big data are combined. This is because machine learning techniques are becoming more widely recognized across domains. There is an urgent need to solve this issue, and given current developments and trends, it is possible that new problems and creative solutions will arise [24].

An outline of class imbalance categorization and the ensuing difficulties is provided in this work [25]. It explains the primary obstacles that impede the classifier's ability to handle extremely unbalanced datasets as well as the numerous variables that lead to class imbalance concerns [26]. Justifications for this study attempt are discussed along with research gaps in earlier publications. Finally, the current patterns that have been identified are discussed along with the latest developments in the classification of class imbalance. Additionally, this study makes various recommendations for future directions in the field, including the use of machine learning in big data computing and the explosion of sentiment analysis from social media.

8. CONCLUSION

An outline of class imbalance categorization and the ensuing difficulties is provided in this work. It explains the primary obstacles that impede the classifier's ability to handle extremely unbalanced datasets as well as the numerous variables that lead to class imbalance concerns. Justifications for this study attempt are discussed along with research gaps in earlier publications. Finally, the current patterns that have been identified are discussed along with the latest developments in the classification of class imbalance. Additionally, this study makes various recommendations for future directions in the field, including the use of machine learning in big data computing and the explosion of sentiment analysis from social media.

References

1. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer , "SMOTE: Synthetic Minority Over-sampling Technique, 2002
2. Edwardo A. Garcia "Learning from Imbalanced Data",2009
3. A. Fernández, S. García, M. J. del Jesus, and F. Herrera, "Random Forests for Imbalanced Data", 2013
4. Pablo M. Olmos and Emilio Soria-Olivas, "Addressing Imbalanced Classification Problems by Combining SMOTE and Ensemble Classifiers" ,2014
5. Longbing Cao, Chengqi Zhang, and Guandong Xu, "Cost-sensitive Boosting for Classification of Imbalanced Data", 2013
6. Michele Filannino, Marco Guerini, and Alberto Lavelli, "Addressing Data Sparsity and Class Imbalance in Online News Popularity Prediction" , 2015
7. American Diabetes Association, ''Diagnosis and classification of diabetes mellitus,'' Diabetes Care, vol. 35, pp. S64–S71, Jan. 2012.
8. V. Jaiswal, A. Negi, and T. Pal, ''A review on current advances in machine learning based diabetes prediction,'' Primary Care Diabetes, vol. 15, no. 3, pp. 435–443, Jun. 2021
9. P. Kaur, N. Sharma, A. Singh, and B. Gill, ''CI-DPF: A cloud IoT based framework for diabetes prediction,'' in Proc. IEEE 9th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON), Nov. 2018, pp. 654–660.
10. S. Chatterjee, K. Khunti, and M. J. Davies, ''Type 2 diabetes,'' Lancet, vol. 389, no. 10085, pp. 2239–2251, Jun. 2017.
11. G. S. Aujla, A. Jindal, R. Chaudhary, N. Kumar, S. Vashist, N. Sharma, and M. S. Obaidat, ''DLRS: Deep learning-based recommender system for smart healthcare ecosystem,'' in Proc. IEEE Int. Conf. Commun. (ICC), May 2019, pp. 1–6.
12. Y. LeCun, Y. Bengio, and G. E. Hinton, ''Deep learning,'' Nature, vol. 521, no. 7553, pp. 436–444, Dec. 2015
13. D. S. J. Ting, V. H. Foo, L. W. Y. Yang, J. T. Sia, M. Ang, H. Lin, J. Chodosh, J. S. Mehta, and D. S. W. Ting, ''Artificial intelligence for anterior segment diseases: Emerging applications in ophthalmology,'' Brit. J. Ophthalmol., vol. 105, no. 2, pp. 158–168, Feb. 2021.
14. S. Nijman, A. Leeuwenberg, I. Beekers, I. Verkouter, J. Jacobs, M. Bots, F. Asselbergs, K. Moons, and T. Debray, ''Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review,'' J. Clin. Epidemiol., vol. 142, pp. 218–229, Feb. 2022
15. M. Alabadla, F. Sidi, I. Ishak, H. Ibrahim, L. S. Affendey, Z. C. Ani, M. A. Jabar, U. A. Bukar, N. K. Devaraj, A. S. Muda, A. Tharek, N. Omar, and M. I. M. Jaya, ''Systematic review of using machine learning in imputing missing values,'' IEEE Access, vol. 10, pp. 44483–44502, 2022
16. J. M. Johnson and T. M. Khoshgoftaar, ''Survey on deep learning with class imbalance,'' J. Big Data, vol. 6, no. 1, pp. 1–54, Mar. 2019.
17. D. Solomon, S. Patil, and P. Agrawal, ''Predicting performance and poten- tial difficulties of university student using classification: Survey paper,'' Int. J. Pure Appl. Math, vol. 118, no. 18, pp. 2703–2707, 2018.
18. E. Alyahyan and D. Düştegör, ''Predicting academic success in higher education: Literature review and best practices,'' Int. J. Educ. Technol. Higher Educ., vol. 17, no. 1, Dec. 2020.
19. V. L. Miguéis, A. Freitas, P. J. V. Garcia, and A. Silva, ''Early segmen- tation of students according to their academic performance: A predictive modelling approach,'' Decis. Support Syst., vol. 115, pp. 36–51, Nov. 2018.
20. P. M. Moreno-Marcos, T.-C. Pong, P. J. Munoz-Merino, and C. D. Kloos,''Analysis of the factors influencing Learners' performance prediction with learning analytics,'' IEEE Access, vol. 8, pp. 5264–5282, 2020.
21. A. E. Tatar and D. Düştegör, ''Prediction of academic performance at undergraduate graduation:

Course grades or grade point average?'' Appl. Sci., vol. 10, no. 14, pp. 1–15, 2020.
22. Y. Zhang, Y. Yun, H. Dai, J. Cui, and X. Shang, ''Graphs regularized robust matrix factorization and its application on student grade prediction,'' Appl. Sci., vol. 10, p. 1755, Jan. 2020.
23. H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, ''Educational data mining and learning analytics for 21st century higher education: A review and synthesis,'' Telematics Informat., vol. 37, pp. 13–49, Apr. 2019.
24. K. L.-M. Ang, F. L. Ge, and K. P. Seng, ''Big educational data & analytics: Survey, architecture and challenges,'' IEEE Access, vol. 8, pp. 116392–116414, 2020.
25. A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, ''Predict- ing academic performance: A systematic literature review,'' in Proc. 23rd Annu. Conf. Innov. Technol. Comput. Sci. Educ., Jul. 2018, pp. 175–199.
26. L. M. Abu Zohair,''Prediction of student's performance by modelling small dataset size,'' Int. J. Educ. Technol. Higher Educ,vol. 16, no. 1, pp. 1–8, Dec. 2019, doi: 10.1186/s41239-019-0160-3.