Online-Journals.org

iTDAF IETI Transactions on Data Analysis and Forecasting

iTDAF | eISSN: 2959-0442 | Vol. 2 No. 3 (2024) | OPEN ACCESS

https://doi.org/10.3991/itdaf.v2i3.51013

SHORT PAPER

Improving the Imbalanced Data Accuracy Using CNN and ReLU

Adnan Saeed[1], Junaid Baber[1], Muhammad Zain Abbas[1], Ahthasham Sajid[2](✉), Hamza Razzaq[2], Arslan Ali Khan[2]

[1]Department of Computer Science and Information Technology, University of Baluchistan, Quetta, Pakistan

[2]Department of Cyber Security, Riphah Institute of Systems Engineering, Riphah International University, Islamabad, Pakistan

ahthasham.sajid@riphah.edu.pk

ABSTRACT

In today's academic world, learning from datasets is a trendy topic. In the event of imbalanced data, none of the many data mining tools available for this purpose work well, partly because this type of data generates a range of minority classes, which might obstruct the learning process. In addition to its enormous volume, big data has the traits of speed and variety. In this paper, the authors have proposed a CNN model with a rectified linear activation function (ReLU) activation function to get good accuracy on an imbalanced dataset. The dataset used in this paper was electrocardiogram (ECG) heartbeat categorization.

KEYWORDS

electrocardiogram (ECG), DLIB, CNN, rectified linear activation function (ReLU), TP, TN, FP, FN, RDBI, generalized Lloyd algorithm (GLA)

1 INTRODUCTION

Researchers are currently focusing on extracting knowledge from large datasets generated in domains such as healthcare, financial services, and telecommunication. The issue with these real-world datasets is that they are frequently imbalanced. The purpose of classification predictive modeling is to predict a class label for a given observation. The distribution of examples across recognized classes is unequal or biased in an imbalanced classification task. The imbalance might range from a little skew to hundreds, thousands, or millions of cases in the majority class or classes. Because most machine learning techniques for classification are based on the premise that each class has an equal number of samples, uneven classifications make predictive modeling challenging. As a result, models with low prediction accuracy evolve, especially for the minority group [10]. For example, a medical diagnosis dataset that contains samples that match the diagnosis of an abnormal condition, but only 5% of the sample matches a positive diagnosis. The benign group, in other words, makes up the bulk of the remaining samples. Machine learning systems designed to perform well on balanced datasets struggle when learning from unbalanced datasets. When

Saeed, A., Baber, J., Abbas, M.Z., Sajid, A., Razzaq, H., Ali Khan, A. (2024). Improving the Imbalanced Data Accuracy Using CNN and ReLU. *IETI Transactions on Data Analysis and Forecasting (iTDAF)*, 2(3), pp. 50–58. https://doi.org/10.3991/itdaf.v2i3.51013

Article submitted 2024-07-09. Revision uploaded 2024-08-26. Final acceptance 2024-08-26.

learning from these unbalanced datasets, machine learning methods that were designed to perform well with balanced datasets are impeded. The problem persists since these algorithms try to increase accuracy in the majority of circumstances [11].

1.1 Rectified linear activation function

In a neural network, the activation function oversees transforming the node's summed weighted input into the node's activation or output for that input. The rectified linear activation function, or ReLU for short, is a piecewise linear function that outputs the input directly if the input is positive and 0 otherwise. It has become the default activation function for many types of neural networks since it is faster to train and generally generates better performance. The layers of nodes in a neural network learn to map different sorts of inputs to distinct types of outputs. The inputs are multiplied by a node's weights, then combined together for that node. This value represents the total activation of the node. The aggregate of the activations is then changed using an activation function, which determines the individual output or "activation" of each node.

2 LITERATURE REVIEW

By means of cluster analysis, data were divided into different groups based on natural features of similarity, measurement, and perception, thus identifying and underestimating the amount of data. Cluster analysis has been widely used in many fields, such as image segmentation, text mining, bioinformatics, wireless sensor networks, and financial analysis [1].

Conventional clustering algorithms can be divided into two categories: the partition method separates a sample of data into multiple clusters simultaneously, so that each event is for a specific group only. They are used because they are simple, efficient, and easy to use. Hierarchical methods and techniques form a class with collective categories in combination or in a split mode. KM clustering must face some limitations: initialization sensitivity, noise inclination, and exposure to unwanted pattern distribution [2].

When a minority is wrongly recognized as the majority, you wind up paying a lot of money for a lot of recordings that are incorrectly priced and have low racial recognition. Improving the degree of a few classes has become a key challenge in machine learning and pattern recognition. The current state of research on the problem of unbalanced data set classification is as follows: Evaluation criteria, as well as a solution When we attempt to address all difficulties, we may split them into two categories. As a result, the major focus of research into the separation of unequal data sets is on how to enhance the subdivision characteristics, which is one of the two separation difficulties. These are two examples of TP and TN. The number of tiny or small samples in the minority category and the general category should both be included in the FN and FP categories (if the category is not yet well divided). Some of the machine learning performance evaluation parameters i.e. (precision, G means, recall, clarification, and F value). Small partitions have the ability to address the challenge of dividing unequal data sets. Algorithm-level approaches and data-level methods are separated. In a collection, data or items are collected by distinct groups based on some of their features, with the goal of making the data or objects in the group as similar as possible while increasing the variety of the different groups. The k-mean algorithm is a distance-based technique that utilizes distance as a measure of similarity. The shorter the distance between two items, the closer they are. This is the distance between the data point and the type (separation center) as well as the desired goal function. Find the guidelines for repeat modifications using the overall technique at work [3].

A set is a learning area in which one course has more examples than the other. The problem with standard data sets is that the classification learning methods are usually appropriate for most categories (known as the "Negative" category), resulting in a high level of sub-category misclassification (so-called "positive" examples). A "class imbalance" is the term used to describe this situation. Finding rare diseases, unreliable communication customers, oil spills on satellite radar images, learning to pronounce language, and other decreases in specimens from tiny races are all costly. text separation and long-distance calls details and filters for your search. The following are some possible solutions to this issue: A data layer, which includes under-sampling, over-sampling, and other data-based enhancement approaches. Like a conventional segmentation, an uneven section distribution can be used as a segmentation divider. Algorithm level, for example, cost-sensitive learning, one-sided learning, and so on. The strategy focuses on improving fundamental teaching practices in order to better address the issue of class imbalance. Change possible limitations on tree leaves (if using decision tree), adjust decision limit, and make acquisition (i.e., read in section) cost of various categories to offset category disparity. As a result of the poor categorization of excellent and bad class models, greater expenses result. As a result, strive to decrease the number of expert hybrid approaches (hybrid techniques) employed to address the issue of class disparity.

To tackle the problem of intermediate inequality and, this paper combines the sampling oversampling method with the K-means-based sampling algorithm [4]. The rough k-mean clustering technique and its expansion have been used to successfully assemble actual data with no particular parameters. The "rough k mean clustering algorithm" offers a binding set above and below the provided data set, according to tests. The same amount of weight When a new center is calculated for each group, it applies to virtually all data items in the next or preceding group, and the dynamics of distinct objects in the same zoom are disregarded. Grouping is one of the most frequent data analysis techniques that may be used to nearly any issue. The clustering algorithm divides data items into groups in order to make things the same within and across groupings. The grouping techniques may be separated into division method, density method, hierarchy method, grid method, and model method, according to the findings of this study.

The authors in this paper explain how to utilize "k-means" effectively by first calculating the "kd tree" in a collection of data. Second, the "kd tree" is used to select the first centers of clusters found in densely populated and well-separated areas; third, to update cluster centers by repeatedly accessing the nodes in the "kd tree" and without including the distance calculation in the cluster center at least one point in the first place; and finally, to update cluster centers by repeatedly accessing the nodes in the "kd tree" and without including the distance calculation in the cluster center at least one point in the first place [5].

The performance of k-means can be enhanced by decreasing the distance computation done at the collecting center near the point distance to the collection center that can be discovered at least one point in the node is determined from the data points in the "kd tree." The missing collection center can be shortened by calculating the lower and higher distances than the region. The "Kd tree" is a data structure for organizing data points in a space by dividing space. In the top-level split scheme, represent data points. All points are in a rectangle. The "kd tree" is where the "K-means algorithm" stores input data points. As the first collection center, choose the first random point from the data set. The cluster center can be updated by calling all nodes in the "kd tree" in the sequence structure from root node to leaf node. Each node has several student centers, which serve as the nearest centers and serve as the foundation for cluster centers. Neighborly nodes that have at least one point in common can be classified as "Complete node, Effective node, or Ineffective node" based on the size of their baptismal centers. The execution time was decreased by 31% to 35%, while the distance computation time was reduced by 33% to 60%. In comparison to contemporary implementation approaches, experimental testing of comprehensive and accurate data sets demonstrates that the RDBI can cut filtering time by 5%. Between 2% and 15% of the time, the accuracy of the group findings improves by 35 percent. Data clustering splits items into randomly separated groups in order to improve object "intra-cluster similarity" while reducing object "inter-cluster similarity" [6].

Cluster analysis is useful in a variety of activities, including "vector quantization and data compression," "pattern recognition," "data and information discovery," "clustering of similar gene expression," "document clustering," and "image processing" (from top to bottom). Tree structure represents the many cutting planes that are used to create clusters. To discover groups, multi-functional collecting methods employ quantitative data from surrounding items. Lloyd presented a technique of multiplication based on the basic observation that the facility is focused on the data set's group. The "Generalized Lloyd Algorithm (GLA)" is the name given to it. Lloyd's extended version is known as the k-means clustering method, and it recognizes scalar data and adds it to vector data. "Fuzzy C-means" is a "k-means clustering" extension. For a particular degree of membership, each point is assigned to several sets. This paper explains how to utilize "k-means" effectively by first calculating the "kd tree" in a collection of data.

The technique of "fault categorization" is an important element of the monitoring of the manufacturing process [7]. Many classic "fault classification" approaches presume that the amount of data collected at each stage is equal. The bulk of data from industry operations, however, is traditional data (majority), with just a few incorrect (minority) data. As industrial systems become increasingly sophisticated and interconnected, process monitoring becomes the most critical component in maintaining the process's dependability enhance the quality of the product large amounts of process data were captured and stored as information technology progressed. As a result, data-driven techniques are fast evolving and becoming more industrialized.

A scenario in which the number of events in at least one category is larger than the number of instances in other classes is referred to as "class imbalance." For the majority or minority, this position is problematic. Because it may exhibit unique behavior that varies from regular access patterns, the minority group is extremely worried. Furthermore, because the problem of "class imbalance" tends to transfer to the dominant class, it causes a distortion in the decision-making process [8].

The goal of this study is to use "K-means" to identify the collection's centroid. The influence of the group process on class disparity is also examined. The perceptron feed-forward neural network is recommended. Complicated pattern recognition is a technique for detecting complex patterns. The algorithm falls from simplicity into a more complicated autonomic neurogenic learning machine with a flexible neural network shared as part of it. To enhance the learning level of the two-layer imbalance problem, more study utilizing the neural network approach of "modified distribution" was recommended. The most used clustering method is the "K-means algorithm." It provides an excellent clustering approach for large numerical data, allowing the same data to be placed in the same clusters. A "tri-level, bi-layer k-means method" is used in this work. Outliers, acoustic data, and initial clusters all use the "k-means method." These flaws are solved by the "tri k-means" method. Pattern recognition, picture extraction, data mining, and data compression are all used in numerous locations. Combining data in the same cluster with different information in a separate cluster is an excellent approach to doing so. Without tampering with the unknown data, the clustering process separates the data into a certain group [9].

3 RESULTS AND DISCUSSION

We have used the famous electrocardiogram (ECG) Heartbeat Categorization dataset. The MIT-BIH Arrhythmia Dataset and the PTB Diagnostic ECG Database, two well-known datasets for heartbeat classification, were used to create this dataset, which consists of two sets of heartbeat signals.

There are enough samples in both collections to train a deep neural network. This dataset was used to test specific transfer learning skills as well as examine heartbeat categorization using deep neural network architectures. The signals correlate to ECG forms of heartbeats in the normal case and cases affected by different arrhythmias and myocardial infarction.

This dataset has five classes:

- Normal beat
- Unknown beat
- Ventricular ectopic beat
- Supraventricular ectopic beat
- Fusion beat

3.1 Imbalanced dataset

This ECG dataset was an imbalanced dataset, as you can see in Figure 1. Class 1 had more than 82% of the data from the dataset, which shows how inequal the dataset was. Whereas the other remaining four classes were sharing the remaining 18% of the data between each other. Also, class 2, which is Fusion Beats, has the lowest data among others.

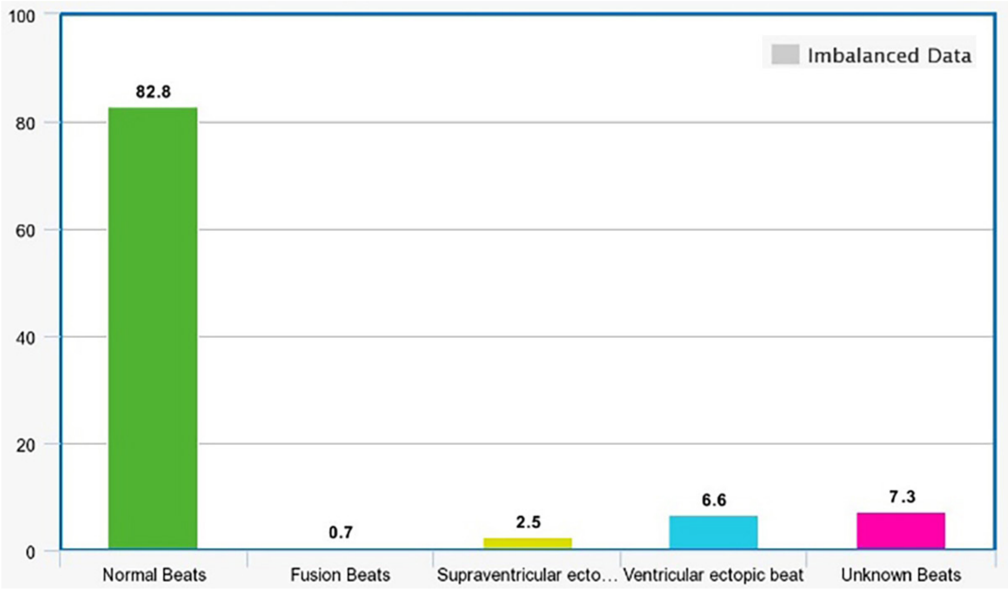

Fig. 1. Imbalanced data

3.2 Resampling for balancing the dataset

As shown in Figure 1, you can see the data was imbalanced. To balance it, we used the famous technique for balancing the data known as the resampling technique. After applying the technique as shown in Figure 2, you can see that now each class has the same data. Because of that, now we can process it further. By processing author means applying the proposed neural model with activation function known as rectified linear activation function.

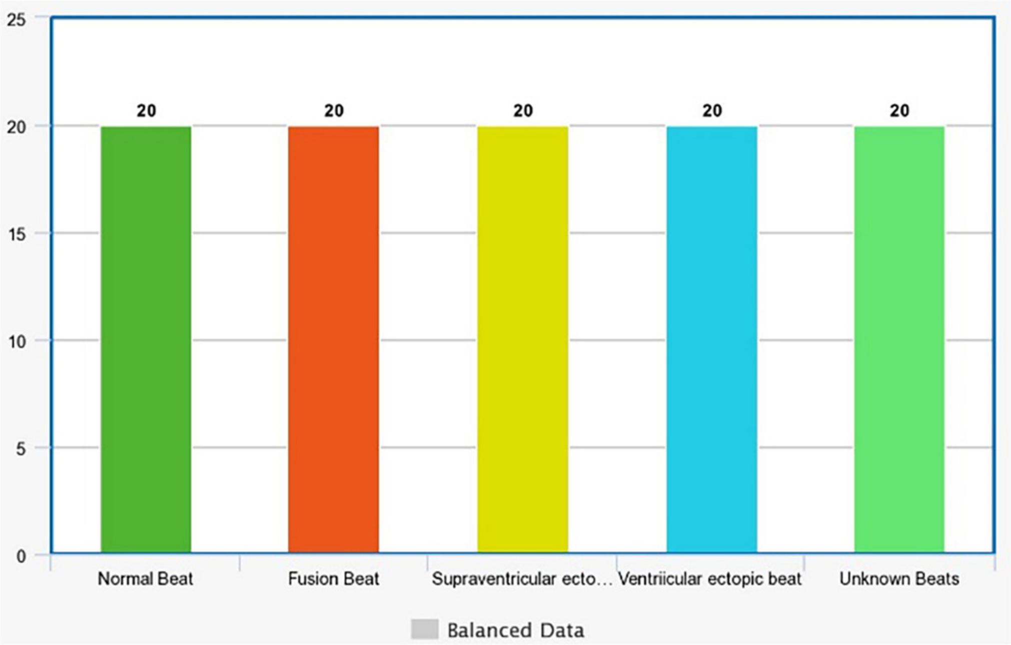

Fig. 2. Balanced data

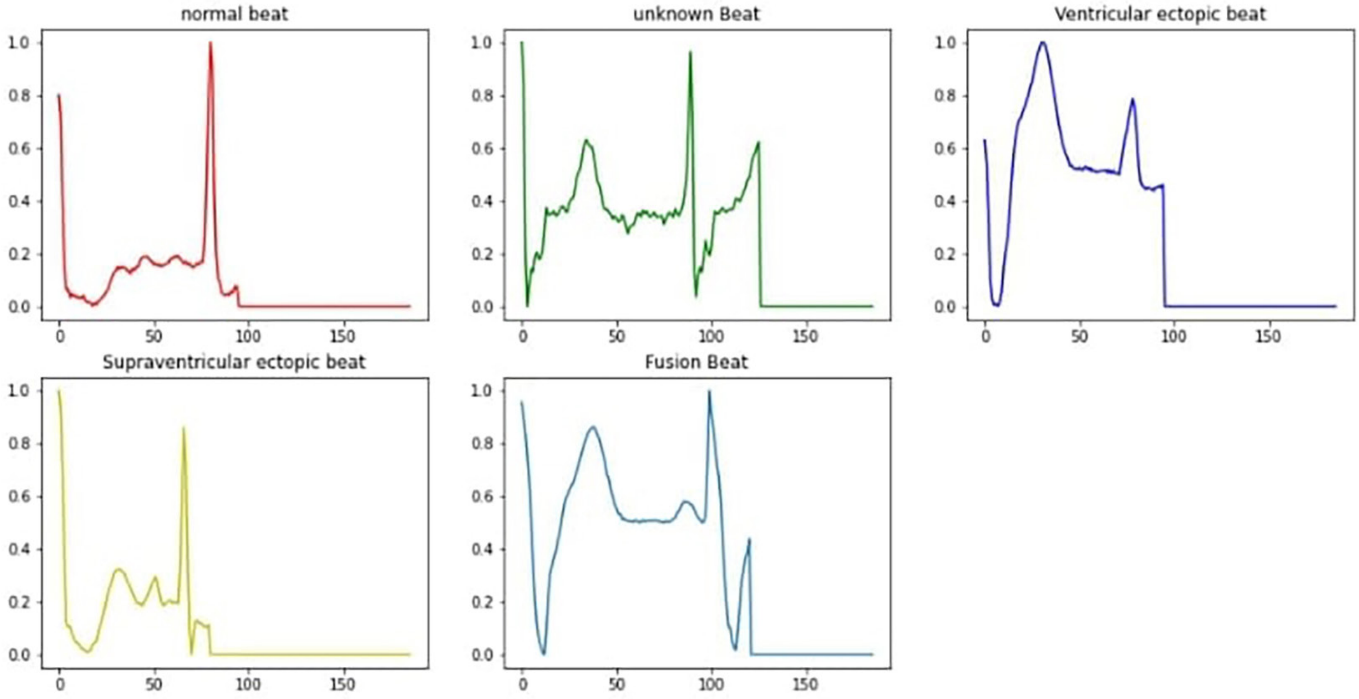

Fig. 3. Visualization of electrocardiogram

The visualization of each class is as follows: this is what each class's ECG looks like. As you can see, each class has a different type of heartbeat, and because of this, every ECG is different from each other, as illustrated in Figure 3. We are creating our own CNN, in which we are using the ReLU. The authors are using the ReLU function because it addresses the problem of disappearing gradients, allowing models to learn more quickly and perform better.

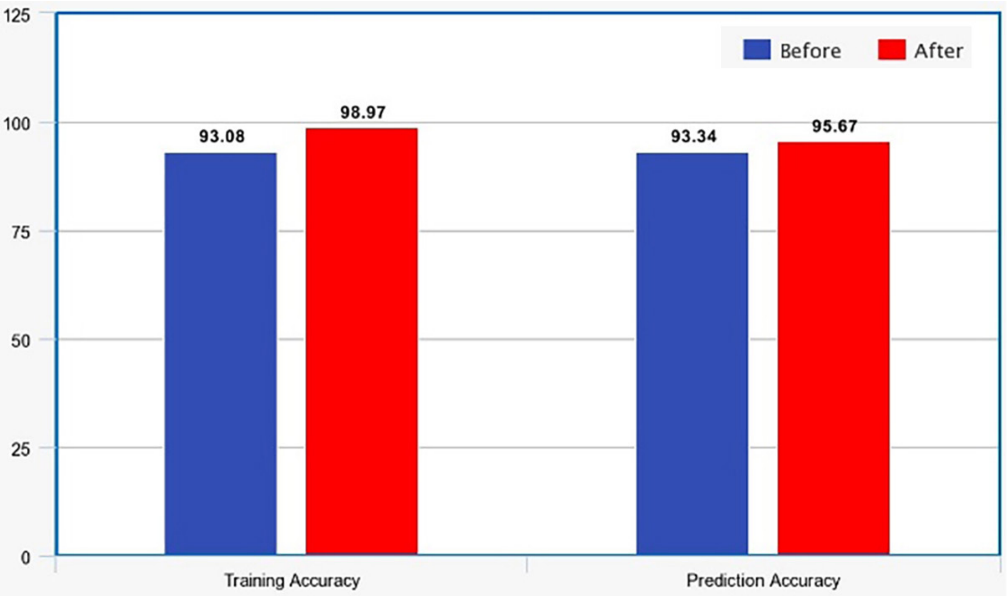

Fig. 4. Pre and post test

Before running our model on the dataset, it was giving training and prediction accuracy around 93% and 93.3% respectively (see Figure 4). After applying the proposed model to the given data, the authors have been able to improve the accuracy of training and predicting data. It was nearly improved, 5% more accurate in both phases.

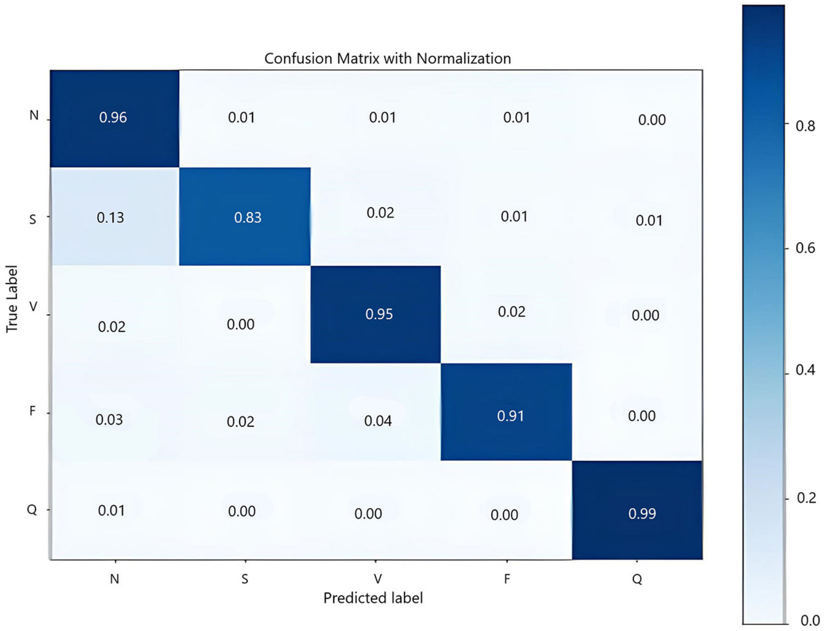

Fig. 5. Confusion matrix

A confusion matrix is a way of describing the performance of a classification system. Classification accuracy alone might be deceiving if the amount of data in each class is unequal. As shown in Figure 5. Therefore, authors have presented the confusion matrix. The confusion matrix with normalization represents five classes with true labels and predicted labels. As you can see in the confusion matrix, our model prediction performance was very good.

4 CONCLUSION AND FUTURE WORK

In recent research, several methods for enhancing the performance of a classifier from imbalanced datasets have been developed, the most prominent of which is the use of re-sampling algorithms, which learn from the data by adjusting its categories in favor of a specific viewpoint. By using the resampling technique on the CNN model with the help of the ReLU activation function, great results can be achieved. The authors of the study have improved 5% accuracy through the proposed model.

In the future the accuracy of the imbalance dataset can be enhanced using deep learning algorithms, i.e., CCN, LSTM, etc.

5 REFERENCES

[1] H. Xie *et al*., “Improving k-means clustering with enhanced firefly algorithms,” *Applied Soft Computing*, vol. 84, p. 105763, 2019. https://doi.org/10.1016/j.asoc.2019.105763

[2] Y. Yong, “The research of imbalanced data set of sample sampling method based on k-means cluster and genetic algorithm,” *Energy Procedia*, vol. 17, pp. 164–170, 2012. https://doi.org/10.1016/j.egypro.2012.02.078

[3] J. Song, X. Huang, S. Qin, and Q. Song, “A bi-directional sampling based on k-means method for imbalance text classification,” in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 2016, pp. 1–5. https://doi.org/10.1109/ICIS.2016.7550920

[4] T. Zhang and F. Ma, “Improved rough k-means clustering algorithm based on weighted distance measure with Gaussian function,” *International Journal of Computer Mathematics*, vol. 94, no. 4, pp. 663–675, 2016. https://doi.org/10.1080/00207160.2015.1124099

[5] P. Shukla and K. Bhowmick, “To improve classification of imbalanced datasets,” in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017, pp. 1–5. https://doi.org/10.1109/ICIIECS.2017.8276044

[6] K. M. Kumar and A. R. M. Reddy, “An efficient k-means clustering filtering algorithm using density based initial cluster centers,” *Information Sciences*, vols. 418–419, pp. 286–301, 2017. https://doi.org/10.1016/j.ins.2017.07.036

[7] G. Chen, Y. Liu, and Z. Ge, “K-means Bayes algorithm for imbalanced fault classification and big data application,” *Journal of Process Control*, vol. 81, pp. 54–64, 2019. https://doi.org/10.1016/j.jprocont.2019.06.011

[8] Hartono *et al*., “Optimization model of k-means clustering using artificial neural networks to handle class imbalance problem,” in *IOP Conference Series: Materials Science and Engineering*, vol. 288, 2018, no. 1, pp. 1–9. https://doi.org/10.1088/1757-899X/288/1/012075

[9] S. S. Yu, S. W. Chu, C. M. Wang, Y. K. Chan, and T. C. Chang, “Two improved k-means algorithms,” *Applied Soft Computing*, vol. 68, pp. 747–755, 2018. https://doi.org/10.1016/j.asoc.2017.08.032

[10] D. Kibler and D. W. Aha, “Learning representative exemplars of concepts: An initial case study,” in *Proceedings of the Fourth International Workshop on Machine Learning*, 1987, pp. 24–30. https://doi.org/10.1016/B978-0-934613-41-5.50006-4

[11] F. Provost, “Machine learning from imbalanced data sets 101,” in *Proceedings of the AAAI’2000 Workshop on Imbalanced Data*, 2000.

6 AUTHORS

Adnan Saeed is with the Department of Computer Science and Information Technology, University of Baluchistan, Quetta, Pakistan.

Junaid Baber is with the Department of Computer Science and Information Technology, University of Baluchistan, Quetta, Pakistan.

Muhammad Zain Abbas is with the Department of Computer Science and Information Technology, University of Baluchistan, Quetta, Pakistan.

Ahthasham Sajid is with the Department of Cyber Security, Riphah Institute of Systems Engineering, Riphah International University, Islamabad, Pakistan (E-mail: ahthasham.sajid@riphah.edu.pk).

Hamza Razzaq is with the Department of Cyber Security, Riphah Institute of Systems Engineering, Riphah International University, Islamabad, Pakistan.

Arslan Ali Khan is with the Department of Cyber Security, Riphah Institute of Systems Engineering, Riphah International University, Islamabad, Pakistan.