Global Transitions Proceedings 3 (2022) 91–99

Contents lists available at ScienceDirect

Global Transitions Proceedings

journal homepage: http://www.keaipublishing.com/en/journals/global-transitions-proceedings/

A review: Data pre-processing and data augmentation techniques

Kiran Maharana*, Surajit Mondal, Bhushankumar Nemade

Department of Information Technology, Thakur College of Engineering and Technology, Mumbai University, India

ARTICLE INFO

Keywords:
Data augmentation
Data cleaning
Data oversampling
Data pre-processing
Data wraping

ABSTRACT

This review paper provides an overview of data pre-processing in Machine learning, focusing on all types of problems while building the machine learning problems. It deals with two significant issues in the pre-processing process (i). issues with data and (ii). Steps to follow to do data analysis with its best approach. As raw data are vulnerable to noise, corruption, missing, and inconsistent data, it is necessary to perform pre-processing steps, which is done using classification, clustering, and association and many other pre-processing techniques available. Poor data can primarily affect the accuracy and lead to false prediction, so it is necessary to improve the dataset's quality. So, data pre-processing is the best way to deal with such problems. It makes the knowledge extraction from the data set much easier with cleaning, Integration, transformation, and reduction methods. The issue with Data missing and significant differences in the variety of data always exists as the information is collected through multiple sources and from a real-world application. So, the data augmentation approach generates data for machine learning models. To decrease the dependency on training data and to improve the performance of the machine learning model. This paper discusses flipping, rotating with slight degrees and others to augment the image data and shows how to perform data augmentation methods without distorting the original data.

1. Introduction

Machine learning applications in all technology fields and applied in real-life problems continue to diversify and increase rapidly. The performance of Machine Learning models depends on the quantity, quality, and diversity of data. To enhance the algorithm's reliability, it is essential to select the target data which is to be chosen from the original data set. Data can be acquired in the form of symbolic and numeric attributes, which can be obtained from human beings to sensors to different degrees of complexity and quality of trustworthiness [1].

The facts or figures from which different conclusions can be drawn, or it can interpret and present information from the raw data researcher must discuss the data source [3,4]. A large variety of factors influence the success of a Machine Learning Algorithm; some primary issues can be the representation and the quality of the dataset. Specifically, if the dataset contains redundancy, noise, or unreliable data, it becomes difficult for an algorithm to discover Information and provide better performance [5].

New Artificial Intelligence Techniques are being given to machines, allowing them to perceive and comprehend the visual world better than humans [7]. As a result, computer vision has come a long way in the last few years. Researchers and academics have been driven to develop algorithms for such visual impressions thanks to computer vision. Understanding the scene entails parsing the image into a meaningful segment [10]. Object recognition entails recognizing a specific object in image data. Item detection entails undertaking semantic analysis and classifying the object, and object recognition entails identifying a particular thing in image data [12]. WSN uses sensors, gateway, routers, and others connected to a particular topology to record different geographical location environmental aspects [14,21,22]. In the realm of computer vision, convolutional neural networks (CNN) have accomplished a lot. CNNs are a type of neural network that uses When complex images are considered, CNN preserves the properties of the image data by keeping their spatial and temporal connections [2,17]. However, when it comes to big data, many apps lack access to Data such as medical imaging [19]. Large data sets influence the performance of deep convolutional networks. The model's performance can be improved by adding the image's data. Data augmentation is a series of strategies for enlarging and enhancing the size and shape of an image while maintaining the label [21].

It's a technique for generating new data with various data orientations. Data augmentation solves two concerns for researchers: first, it generates more data from a limited amount of data, and second, it minimizes overfitting [2]. In this paper, investigation for different data augmentation techniques is done.

This paper talks about different tactics based on two categories: data warping and oversampling. Other functional strategies for increasing the

* Corresponding author.
E-mail address: kiran11621@gmail.com (K. Maharana).

https://doi.org/10.1016/j.gltp.2022.04.020

Available online 3 April 2022

size of small data sets, such as transfer learning, pre-training, dropout regularization, and batch normalization, are also briefly discussed.

2. Related works

2.1. Data processing and related works

Data can be available in various forms: Structured tables, unstructured tables, Images, Audio files, Videos. Etc. A machine cannot directly understand the free text, video, or image as it is; it is necessary to convert the given data into 1s and 0s. So, raw data cannot be directly fed to the machine learning model and expects it to get trained.

Data Pre-processing is the First step in Machine learning in which the data gets transformed/Encoded so that it can be brought in such a state that now the machine can quickly go through or parse that Data. In other words, it can also be interpreted as that the model algorithm can promptly analyze the features of data.

Data pre-processing is the most important and influential for generalization performance of a Supervised Machine Learning Algorithm. The quantity of training data grows exponentially concerning the input space dimension. According to an estimate, time spent on pre-processing can take up to 50% to 80% of the entire classification process, proving the importance of pre-processing in building a model [4]. It is also essential to improve the data quality for better performance.

Data processing consists of steps to be followed before starting the analysis with the actual data for the model. It is essential that a transformation T that transforms the raw data vector A_{ik} to a set of new data called B_{ij}.

$B_{ij} = T(A_{ik})$

Such that:

i Bij preserves the essential Information in Aik
ii Bij eliminates at least one of the problems of the vector Aik
iii After the process, Bij is more important and valuable than Aik

Here from the above relation:

$i = 1, 2, 3, \ldots$, a where a = number of objects
$j = 1, 2, 3, \ldots$, b where b = number of features extracted after pre-processing
$k = 1, 2, 3, \ldots$, c where c = number of attributes/features available before pre-processing simply, it should imply $j \neq i$. [1]

Data processing is to be carried out for the following reasons:

i Solving problems related to Data may lead to inconsistency and prevent from performing any data analysis.
ii For building the model, it is necessary to understand the features and nature of the data for meaningful analysis.
iii Extracting more needful and meaningful Information from the given set of data.

2.2. Problems with the data

The data collected from any source persists incomplete, noisy and inconsistent data, leading to problems with the data analysis. So, it is needed to rectify the issues beforehand, and they can be classified into three groups – too much Data, too little data, and fractured data. Fig. 1 explains the problems with data in tabular format.

2.2.1. Too much data

In the case of medicine, telecommunications, or space, the volume and velocity of data are too large. It plays a vital role in the limiting factor in performing analysis with a real-time dataset. Other than this, corrupt data can lease weaken the model's predictive capability. Pre-processing the data for proper interpretation is a form feature that condition the input data to allow easier subsequent feature extraction and increased resolution [6]. Reducing the dimension of the data set may help improve the model's performance. It is also essential to pay proper attention to the data set consisting of numerical or any symbolical parameters as it may increase complexity for the model.

2.2.2. Too little data

If the available Data does not include a sufficient amount of data of all kinds, then the reliability of the knowledge gained from the data may be incompetent. Missing attributes may hinder the accuracy of the model. In the case of decision-tree induction, missing attributes may lead to unequal length. At the same time, they are splitting the dataset into training and testing sets. It may lead to unequal distribution of features. If the data used has more than 20% of Data missing, it must be eliminated [6].

2.2.3. Fractured data

Data incompatibility becomes a significant issue if collected from several groups or different platforms. The goal, depth and standard for maintaining and managing the data may vary as the need. Also, the level of details at which the data are stored in the database can differ, leading to problems during the modeling.

2.3. Features in machine learning

A dataset is a collection of data objects referred to as points, patterns, events, cases, samples, observations, or entities [8]. As a result, these data items are frequently characterized by several features that provides with the essential qualities of an entity, such as an object's mass, the time at which an event occurred, and so on. A feature might be an individual measured quality or a portion of a phenomenon that occurred. It can be broadly classified into two types:

1 Categorical: A categorical feature is one whose values are selected from a collection of pre-defined possibilities. For instance, a month's name. Another example is a Boolean set, which contains values such as True and False.
2 Numerical: Characteristics whose values are continuous or integer. They are primarily represented by numbers, with which they share many of their characteristics. For example, how many steps you take every day or how fast you drive your automobile.

2.3. Pre-processing steps

Fig. 2 represents data pre-processing steps and they are discussed in details in the following -

C.1. Data cleaning process

Data cleansing is a process for detecting incorrect or noisy data and correcting them or removing them from the dataset. In general, it works on identifying and replacing incomplete, inaccurate, irrelevant or any other noise data and records. While the techniques used vary according to the demand of the model, the basic steps followed are a. Remove irrelevant or duplicate data

It often happens in the dataset. When data are combined from a different source, scrape data or data from multiple clients. There opportunity for creating duplicate data generate. b. Structural error fixing

The inconsistencies are generated due to mislabeling of categories or classes. It may also develop due to strange naming conventions, typos, or incorrect capitalization. c. Missing values

It is usual to have missing values of particular columns in a dataset. The issue can be generated due to data validation rules or data collection. But it is necessary to consider missing values because it may eliminate the feature of a model due to missing values. If a reasonable number of values are missing, then simple interpolations methods can fill such matters. The most common method used for dealing with it is using mean, median or mode values concerning model features. a. Inconsistent values

It may be due to human error or generated while working with primary data. So it becomes necessary for having a data assessment process

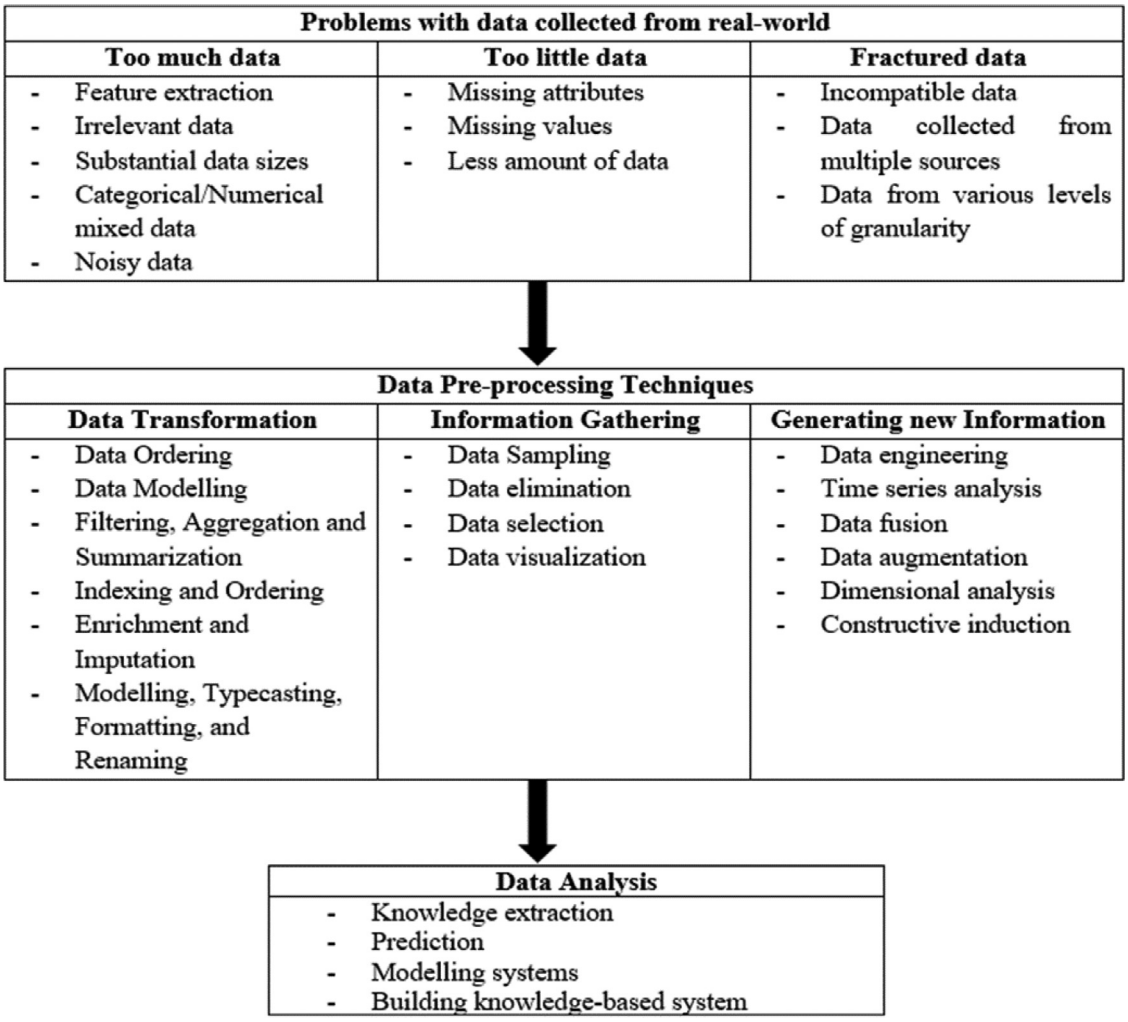

Fig. 1. Problems with data collected from real-world.

to learn the datatype of the feature and check that all data objects are of the same type. b. Data validation

The final dataset should answer the following questions after the process:

- Do these data make sense?
- Does the data adhere to the field's specific rules?
- Does it substantiate or refute the model feature?
- Is it able to identify patterns in the data?

Inconsistent data might lead to erroneous conclusions and forecasts. As a result, high-quality data must meet the following criteria: Validity (Constraints, Range, Patterns, and others), Accuracy, Completeness, Consistency, and Uniformity.

C.2. Noise handling

If the noise continues in class after the loud occurrences have been identified, there are three approaches for dealing with it. First, noise can be ignored if the model is robust enough to tolerate over-fitting. Second, noise in the dataset can be filtered away, modified, polished, or relabeled. If the data with Attribute persists, techniques like filtering or polishing the erroneous attribute value, removing it from the dataset, or imputation can predict what needs to be cleaned and uncover more suspicious values [6]. a. Binning

It's a technique for reducing the impact of minor observation errors. Values are separated into small bins in the original data and then replaced with general values derived for that bin. It smoothes the input data and, in the event of a short dataset, it may lower the chances of overfitting [7,8].

It has two methods –

i Equal Frequency Binning: Binds have an identical frequency.
ii Equal Width Binning: Bins have equal width with range of each bin calculated as [min+w], {min+2w], …, [min+nw] where w = (max-min) / (number of bins)

b. Regression

It is a supervised machine learning technique that is used for the prediction of continuous. It establishes a relation between the variables by estimating how variable affects the other. To evaluate the predictions by regression algorithm, variance and bias metrics must be considered.

i Variance – It is the amount by which the estimate of the target function changed if the training data were different. For avoiding false predictions, the variance of the model should be low.
ii Bias – The algorithm tends to consistently learn the wrong things by not accounting for all the Information of the data. If there are inconsistencies in the dataset like missing values, fewer attributes or errors, it can lead to a biased output. So, it needed to keep the model's bias low to get accuracy.

The most commonly used regression analysis is polynomial regression, linear regression, decision tree, robust regression, Gaussian process regression and support vector regression. c. Clustering

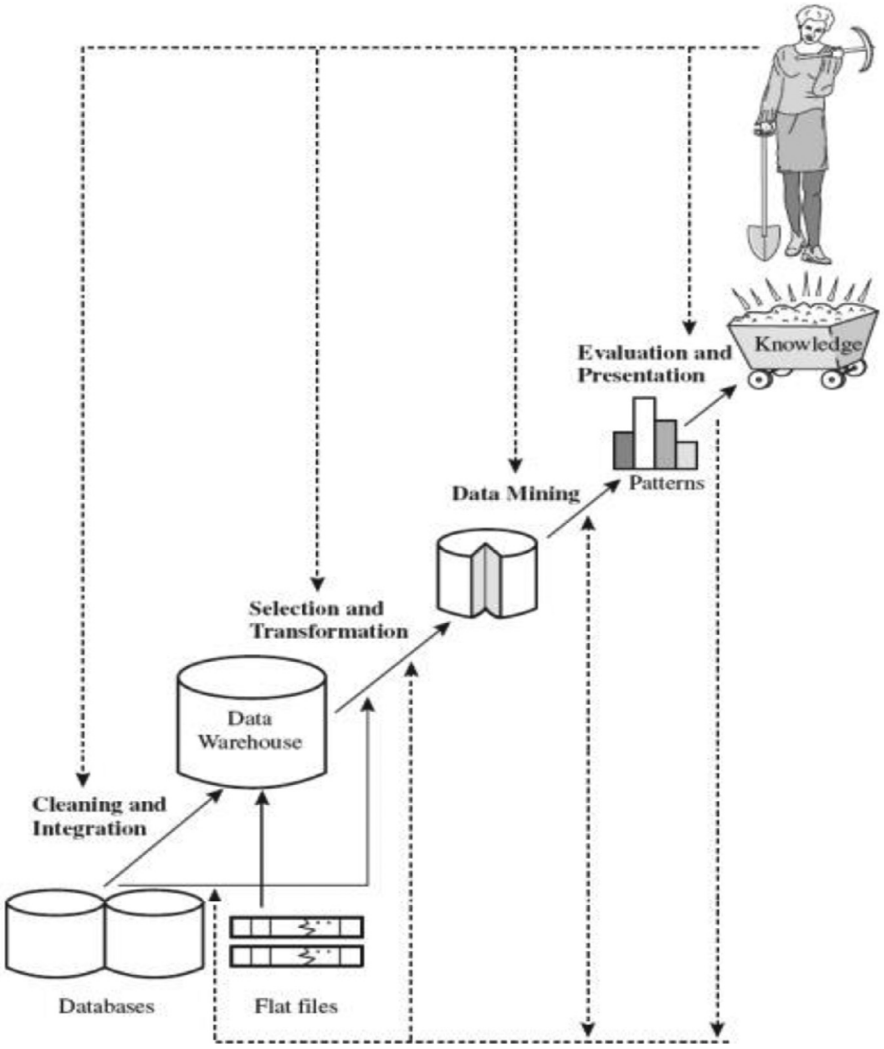

Fig. 2. Data pre-processing steps.

It is an unsupervised learning method. It is a task of dividing the population or data points into clusters (groups). Data points in the same groups are more similar to other issues in the same groups and dissimilar to any other data points.

Clustering is crucial because it determines the intrinsic grouping among the unlabeled data present. There are no criteria for deciding the cluster quality, and just it is required to satisfy the need. It has clustering methods like density-based, hierarchical based, partitioning and grid-based.

Commonly used K-means, Mean-shift, Density-based spatial clustering of application with noise, Expectation-Maximization clustering using gaussian mixture models and Agglomerative hierarchical clustering.

C.3. Data integration

Data integration in pre-processing involves combining multiple heterogeneous data sources into a coherent data store and a unified data view. It can be defined as GSM (Global schema, heterogeneous schema source, and mapping between source and global schema). It has mainly two approaches tight coupling approach and the loose coupling approach [11].

The problems with Integration –

- An overview of an information system's architecture;
- The types of data that need to be managed by competent systems;
- Computer hardware and software, as well as the operating system
- User interface data management software (middleware)
- Data semantics, data models, and schemas
- Constraints on business rules and integrity.

C.4. Data transformation a. Normalization

It is required when attributes on a different scale are considered. When multiple features are present, their characteristics may have a separate scale, so normalization is needed to bring them to the same scale or produce poor results. It includes min-max normalization, z-score normalization, decimal scaling. b. Concept hierarchy generation

Using this low-level or raw Data is substituted with high-level concepts in data generalization. c. Smoothing

It is a process that removes noise from the dataset using an algorithm that highlights essential features present in the dataset. It also helps predict the patterns in it, and it also identifies simple changes to predict different trends to serve the pattern-finding process. d. Aggregation

In data mining, aggregation refers to the act of locating, collecting, and presenting data in a summary format for analysis. When collecting data from many sources, it is crucial to acquire essential and reliable Information. It consists of three steps: data collecting, processing, and displaying data. It can be broken down into time aggregation data points for a single resource over a particular time or spatial aggregation data points for a group of resources over a specific time.

1.A.1. Data reduction

It is a process for reducing the original data volume and representing it in a much smaller volume. It ensures the integrity of the Data while reducing the data. When the Data is of significant importance, it is necessary to reduce it. It may become difficult to learn the desired in-

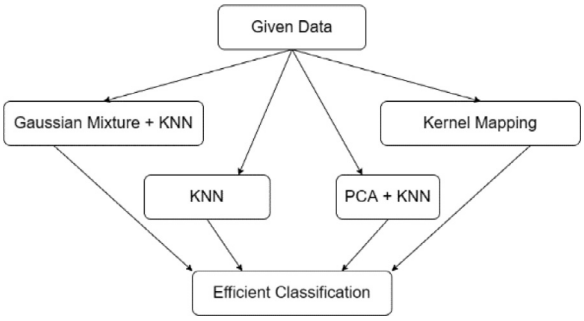

Fig. 3. Data reduction.

formation, and it can take a long time to process complex queries. Fig. 3 explains the process. a. Dimensionality reduction

By deleting properties, dimensionality reduction minimizes the size of the original data collection. The wavelet transforms, principal component analysis or attribute subset selection can reduce dimensionality. b. Numerosity reduction

The numerosity reduction reduces the original data volume, resulting in a significantly smaller representation. There are two sorts of numerosity reduction techniques: parametric and non-parametric.

Parametric - Instead of storing the original Data, parametric numerosity reduction stores only data parameters.

Non-Parametric - Using a histogram to depict frequency distribution, defining how often values appear in the data, grouping, or sampling.

3. Data augmentation and related works

Many scholars worldwide have made progress in this area, identifying several methods for expanding the sample size and, as a result, the generalization of the neural network in use. Using the AlexNet model of CNN Architecture, several authors evaluated and compared various augmentation strategies. ImageNet and CIFAR10 were the datasets used by the authors. Several studies have also reached the efficacy of different augmentation processes such as flipping, rotation, noise, shifting, cropping, PCA jittering GAN, and WGAN. According to the authors, only rotations and WGAN have shown better results than the others. Also, a survey was conducted for semantic segmentation of images and video using deep learning techniques [12].

Using a variety of datasets, the authors explored the complexity of semantic segmentation and evaluated various deep learning models. For picture categorization, a Perlin noise augmentation approach was presented. Different visual patterns are subjected to this pixel-by-pixel Augmentation. They used 106 pictures of patients. They defined 100 regions of interest for each class of graphic ways to be classified using deep learning. When it comes to maximizing the generality of deep understanding, many researchers encounter a significant challenge: the lack of large datasets. Image augmentation, dropout, transfer learning, and other methods are utilized to increase the size of the data set. Innovative Augmentation builds a network that generates augmented Data automatically during the training phase, lowering network loss. The generative adversarial network proposed a method for training synthetic MRI images with brain tumors was proposed by the generative adversarial network [13]. They employed two publicly available brain MRI data sets.

They have highlighted two benefits of using fake data:

i Tumors segmentation with high efficiency.
ii They've shown that generative models can be used anonymously.

A. Data augmentation techniques

Data augmentation is described as a strategy to prevent overfitting via regularization. An intuitive interface has enabled this regularization. To study a task or dataset, one should know what kind of additional Data is needed to enhance the system.

A.1. Symbolic augmentation

These augmentations are classified as "Symbolic Augmentations" instead of "Neural Augmentations." As previously noted, the critical distinction is using ANN, or other forms of statistical models, to generate data instead of symbolic rules. The interpretability for the human creator is a real benefit of symbolic Augmentation. Short transformations, such as substituting words or phrases to produce augmented examples, benefit from symbolic augmentations. However, some information-intensive applications require longer inputs, such as question answering and summarization. Global modifications, such as enhancing entire phrases or paragraphs, are constrained by symbolic restrictions.

A.2. Rule-based augmentation

Rule-based Augmentations create augmented examples by constructing rules. This comprises augmentation programmes and symbolic templates to insert and rearrange existing data.

One of the critical reasons for Easy Data Augmentation's popularity is that it is reasonably simple to use off-the-shelf. Random swapping, deletion, insertion, and synonym replacement are examples of straightforward data augmentation.

A.3. Graph-structured augmentation

For text data augmentation, creating graph-structured representations of text data is an appealing approach. This includes metadata such as citation networks, knowledge graph relation and entity encodings, syntax tree grammatical structures, and metadata underpinning linguistic data. These additions provide Deep Learning systems with comprehensive structural information, which is a very new integration. The addition of structure can help with finding label-preserving transformations, representation analysis, and adding prior knowledge to a dataset or application.

A.4. Mixup augmentation

MixUp Augmentation is creating new examples by combining old samples and blending the labels. MixUp, for instance, may concatenate half of one text sequence with half of another series in the dataset to create a new example. MixUp could be one of the most significant interfaces connecting distant spots and illuminating an interpolation path.

Compared to no regularization or utilizing dropout, their trials found a significant improvement in decreasing overfitting.

A.5. Feature space augmentation

It is used to improve data in the intermediate representation space of DNN. Almost all Deep Neural Networks employ a sequential processing structure, in which incoming Data is turned into distributed representations, which are then used to make task-specific predictions. Feature Space Augmentations use noise to differentiate intermediate features from new data instances. Noise could be generated using adversarial controllers or sampled from uniform or Gaussian distributions.

A.6. Neural augmentation

For the augmentation, ANN generates fresh data. It use a model trained on supervised neural machine translation datasets to translate one language to another and back sample new occurrences, or it employs a model trained on generative language modeling to replace masked out tokens or phrases to generate new data. It also goes through how to use neural style transfer in NLP to translate from one writing style to another or from one semantic characteristic to another, such as formal to casual writing.

Obtaining the correct/accurate Data that is most valuable for research and experiment is a difficult challenge, even though Information is readily available. The data should be diverse enough to be displayed in numerous sizes, positions, colors, and lighting situations for the model to perform better over it. Many data augmentation procedures are applied to deal with the problem of a limited amount of data. These methods will aid in extracting Information from existing data. Data Augmentation based on Data Warping. Detail discussion on techniques used in Image Augmentation is done –

A.6.a. Geometric transformations

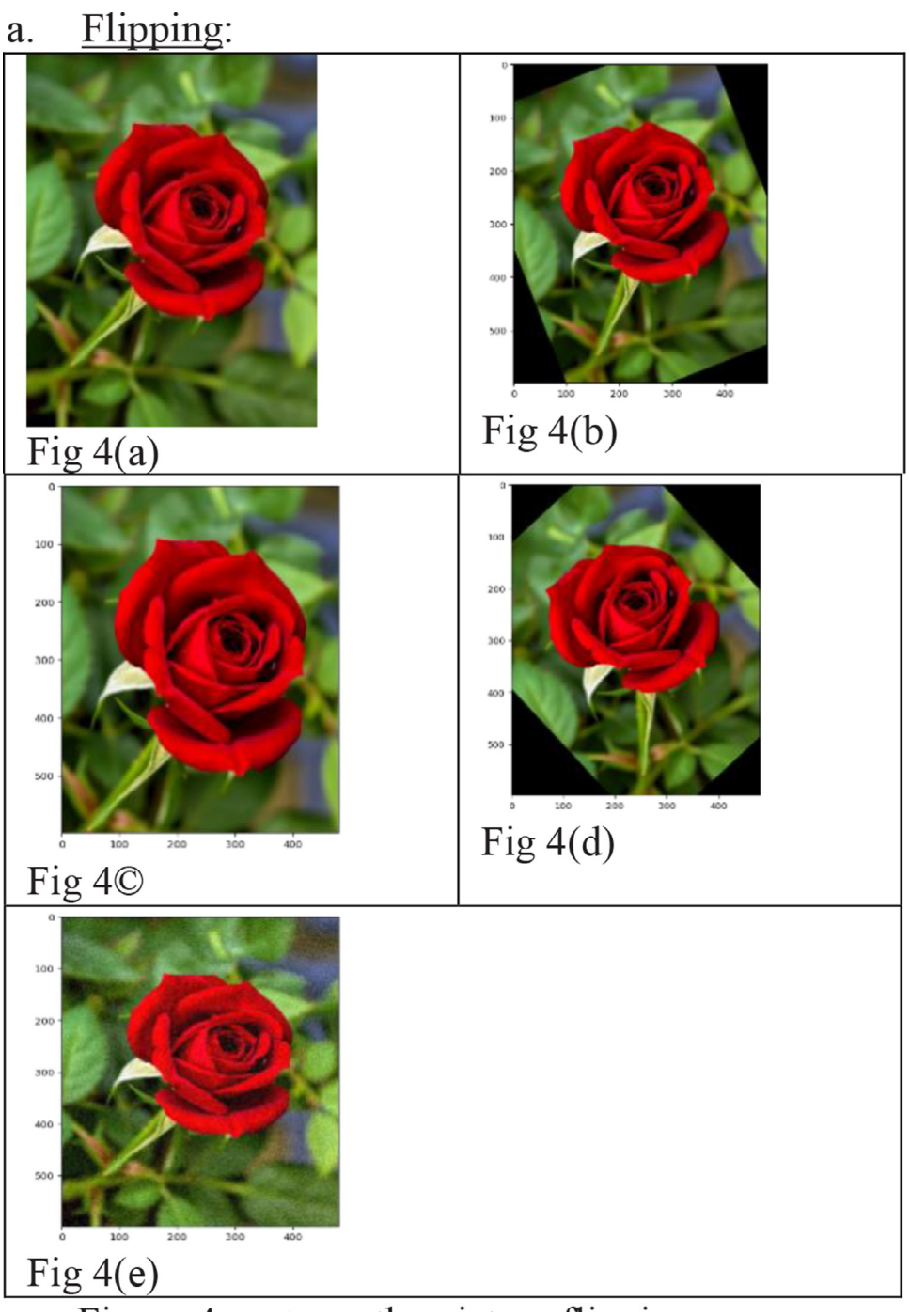

Fig. 4. Portrays the picture flipping process.

This section covers a variety of geometric transformation-based image augmentation techniques and additional image processing techniques. The image is transformed based on its comparable, such as Euclidean, Affine, projectile, etc. [9,13]. The approaches investigated are flipping, color space, cropping, rotation, translation, and noise injection: about different Geometric Segmentation, the ability to preserve the label after transformation is also discussed. a. Flipping

The picture can be flipped on a level plane or in an upward direction in flipping. It makes a picture by pivoting it by a component of ninety degrees. Everything systems do not uphold vertical flipping. Vertical flipping is refined first by pivoting the picture 180°, and afterward level flipping is performed. Fig. 4 portrays the picture flipping process. b. Color space

Color space transformations are known as photometric transformations. A three-photo stacked matrix is created using this technique, with each matrix measuring height X width. This matrix represents the pixel values for each RGB color value. To alleviate the issue of lighting difficulties, the image's color distributions can be modified. c. Cropping

Random cropping is the process of taking a small portion of an original image and resizing it to match the original image's dimensions. If necessary, random cropping is done to resize the image. Translation differs from random cropping because it preserves the image's spatial

Fig. 5. Taxonomy.

dimensions, whereas cropping reduces its size [16]. Fig. 4(c) illustrates how the photo is cropped. d. Rotation

The image can be rotated 90° or positioned at slight angles depending on demands. A photograph is rotated 90° once it has been aligned, and no background noise is inserted. However, this is not the case when turned at a slight angle. The newly added noise will blend in with the rest of the image if the image's background is black or white. The image's backdrop will not blend in entirely if it contains distinct colors, and network will recognize it as part of the image [18]. e. Translation

The translation concept is used to locate the object in any part of the image. Moving the image along the *X* or *Y* direction or both left, right, up, and downshifting images helps avoid positional bias in the data. It aids the network in searching the entire image, resulting in image background noise. f. Noise injection

Noise Injection is a data augmentation approach that inhibits overfitting in neural network models. In most situations, the image is processed with salt and pepper noise. This method requires adding white and black dots to an image, as shown in Fig. 4. (e). Adding noise to the embodiment can help the model learn faster and more effectively [10].

Geometric transformations are a good solution for images with positional biases in training data. More memory and training time and the expense of modification may be drawbacks [11]. Geometric transformations are used in a substantially smaller number of situations.

A.6.b. Color space transformations

CMYK's second multi-dimensional color space translates the first image into the second image. The idea was then divided into three stacked matrices. The height and width of each matrix show its size. Each RGB color value's pixel value is displayed separately. To alleviate the issue of lighting difficulties, the image's color distributions can be modified. To change the photograph's color, you can use a variety of photo editing software. The histogram comprises the pixel values of an image's RGB color channel [2]. The histogram is then changed by applying filters to the image's color space properties. Color transformations increase the amount of memory space, transformation cost, and training time. It can, however, result in lossy changes. RGB, LAB, YCrCb, HSV, and other color conversions are performed [15]. a. Random erasing

One feature that affects Convolutional Neural Networks' ability to generalize is occlusion. If the training samples are not visible, i.e., the CNN model will perform well on the testing images if the images have occlusion. It will, however, fail to recognize/identify the item if it is partially occluded. As a result, random erasing, a new data augmentation approach, is presented to solve the problem and improve the CNN Model's generalization capacity [13]. The image is occluded randomly and with an arbitrary big patch or rectangular region with this technique. b. Adversarial training

Negative training is the process of training neural networks with adversarial examples. It's a cheating algorithm that sees a highly similar image and uses it to fool the classifier. According to a study conducted by researchers, machine learning models can be easily misled by marginally altered photographs [19]. These worried instances are known as "adversarial examples" because they are determined during the input optimization process, maximizing the prediction inaccuracy. c. Neural style transfer

The neural network's ability to modify images in the style of another embodiment is created by isolating and recombing the image content and style. It's a strategy for optimizing photos that uses three images as input: the content image, the desired type, and the input image. Following the data collection, these photos are combined to create an input image that resembles the content image but is painted in the style image.

A. Data augmentation based on oversampling a. Mixing images

Sample Pairing, also known as Image Mixing, is a technique for creating a new sample by overlaying one image with another randomly selected from the training data. After it has been combined, the resultant image is utilized for training the classification model. By randomly selecting two photographs from the training set, it can generate N2 new samples from N samples. Developing a more generalized picture mixing approach allows for more improvisation in this procedure. Images are non-linearly incorporated into fresh training instances. Cropping and repairing can also be done at random [20,22,23]. b. Feature-space augmentation

The ability of neural networks to map high-dimensional images into their low-dimensional representations is one of its strengths. Neural networks map ideas to binary classes or n × 1 vectors. Feature space refers to the lower-level pictures contained within the higher-level layers. According to DeVries and Taylor, noise, interpolation, and extrapolation are all examples of feature space augmentation.

B. Other overfitting solutions

There are various other solutions to solve the problem of overfitting apart from data augmentation. These solutions help to increase the overall performance of the model. Some of the answers are discussed below: c. Transfer learning

Traditional data processing and a range of machine learning algorithms trained on labeled and unlabeled data are used to predict data. They need identical distributions as well as enough space. However, there is heterogeneity in the Data because it originates from some sources [23]. As a result, transfer learning has been proposed as a possible solution to these issues. This system applies Information gained while solving one problem to a similar situation. A fundamental concept has been to reduce the problem of overfitting [24,25]. d. Drop out

Dropout could be a regularization approach in which the hidden units are dropped at random throughout the training period. It's been described as a set of tactics for reducing overfitting by employing a single model to imitate a variety of network architectures while randomly removing specific properties throughout each training iteration. As a result, models becomes more stable. e. Batch normalization

Batch normalization is often referred to as a regularization method. It was agreed that the layer activations would not be equalized. After subtracting the batch mean from each activation, the variance is divided. It standardizes the inputs that the layer receives for each collection. f. Pretraining

Transfer learning is similar to this strategy. Both the weights and the network are communicated in transfer learning, but pre-training allows for significance formation over large datasets.

4. Taxonomy of research

Fig. 5 explains the taxonomy. The first step is removing noisy data from the data set by process of filtering or compression. Then, the next phase of normalization data is aligned such that the variety of data does not impact the model's accuracy or increase biases in the prediction. Then data transformation is done for understanding relationships between the data entity. After this step, missing values treated in the following step values lacking data points are removed. In feature engineering, new features are added up or selected from the dataset for increasing the model accuracy. And at last imbalance data treatment is done using augmentation techniques.

5. Discussion

A review of data preparation and data augmentation methodologies is examined in this paper. The primary purpose of Data pre-processing is to provide data of best quality for data mining. Cleaning methods are used to remove unnecessary data remove all the noise from data. To integrate all the available Data in one place data integration is used with data transformation and data reduction techniques. So the conclusion is that Data pre-processing have and effective role in machine learning and artificial intelligence to make our models more accurate.

Extensive testing of various data augmentation techniques has shown that histogram equalization, random translation, and cutoff can enhance target identification accuracy, but the effect is negligible. Other data augmentation approaches have less of an impact on insulator string recognition than Gaussian blur, scaling, and rotation. The generic convolution neural network-based recognition approach, such as faster RCNN, cannot accept target rotations well, necessitating the inclusion of different methods.

6. Conclusion

This paper discussed various pre-processing and Augmentation techniques for improving the performance and outcomes of machine learning designed models. Firstly, multiple problems related to Data was discussed. Data pre-processing is an essential aspect of every machine learning model since the quality and valuable information obtained from it directly impact model's capacity to learn. Data pre-processing take up to 50% to 80% of the entire classification process. Data Preprocessing techniques such as Data Transformation, Information Gathering and Gathering New Information were briefly discussed. Features of Machine Learning such as Categorical, Numerical was listed. Data Preprocessing steps such as Data Cleaning Process, Remove Irrelevant Data, Fixing Structural error, Missing values, Inconsistent values, Data validation were enlisted. Due to significant variability in building operating features and data quality, existing research show that data pretreatment for making operational data cannot be automated. Currently, it's more of a trial-and-error approach that primarily relies on subject expertise and practical activities. More studies should be focused on automating constructing operational data preparation procedures to improve data analysis efficiency.

For increasing the accuracy of the prediction of the models' various techniques of an image, data augmentation was discussed with the concepts of Data Wrapping and Data Oversampling. Multiple data augmentation solutions were discussed to reduce the problem of overfitting in deep learning models. Examines several Data Augmentation approaches to the problem of Deep Learning models overfitting owing to a lack of data. Data Augmentation has a bright future ahead of it. The potential for using search algorithms that combine data warping and oversampling methods is immense. It is known that to get the highest level of accuracy, deep learning models rely on large datasets, so Data augmentation is based on Data Warping in which Geometric transformations, Cropping, Color space, Rotation, Translation and Noise Injection were discussed briefly. Description of Color space Transformations such as Random Erasing, Adversarial Training and Neural Style Transfer was given. The discussion will help increase the generalization degree of designed models and be applied to different image or video data augmentation domains.

So, before building a model, first, make sure they are working on correct data to prevent inconsistency and understand the features of the data by extracting more needful and meaningful information from the data set. While splitting data, keep a check on features distribution as some may be missed, leading to wrong assumptions. If it has more than 20% of Data missing, simply eliminate it even for a robust model and try to answer whether these data make sense, adhere to the rules, and identify the pattern. Keep a proper check on what values need to be cleaned by keeping the variance and bias of the model to the lowest. Understand business rules and integrity for heterogeneous data. While augmenting the dataset, make sure the process is not creating fake data irrelevant to the data set. Use the random erasing, adversarial training, and neural style transfer techniques to use the best overfitting solution.

References

[1] W.M.S. Famili, Data preprocessing and intelligent data analysis, Intell. Data Anal. 1 (1997) 3–23.

[2] B.S. Saini, C. Khosla, Enhancing performance of deep learning models with different data augmentation techniques: a survey, in: Proceedings of the International Conference on Intelligent Engineering and Management (ICIEM), IEEE, 2020 978-1-7281-4097-1/20/$31.00 ©.

[3] P. Jagannathan, S. Rajkumar, J. Frnda, P.B. Divakarachari, P. Subramani, Moving vehicle detection and classification using gaussian mixture model and ensemble deep learning technique, Wirel. Commun. Mob. Comput. 2021 (2021).

[4] A.I. Kadhim, An evaluation of preprocessing techniques for text classification, Int. J. Comput. Sci. Inf. Secur. 16 (6) (2018) 1947–5500 JuneISSN.

[5] R.K. Dash, T.N. Nguyen, K. Cengiz, A. Sharma, Fine-tuned support vector regression model for stock predictions, Neural Comput. Appl. (2021) 1–15.

[6] J.F. Davis, Process data analysis and interpretation, Adv. Chem. Eng. 25 (2022) Copyright 0 Zoo0 by Academic Press All rights of reproduction in any form reserved. w65-2377/00.

[7] B. Davaasambuu, K. Yu, T. Sato, Self-optimization of handover parameters for long-term evolution with dual wireless mobile relay nodes, Future Internet 7 (2) (2015) 196–213.

[8] xxxSustainable Communication Networks and Application, Springer Scince and Business Media LLC, 2022.

[9] S. Gupta, A. Gupta, S. Gupta, Dealing with noise problem in machine learning data-sets: a systematic review, Procedia Comput. Sci. 161 (2019) 466–474 Procedia Computer Science 00 (2019).

[10] G.B. Rajendran, U.M. Kumarasamy, C. Zarro, P.B. Divakarachari, S.L. Ullo, Land-use and land-cover classification using a human group-based particle swarm optimization algorithm with an LSTM Classifier on hybrid pre-processing remote-sensing images, Remote Sens. 12 (24) (2020) 4135.

[11] P. Ziegler, K.R. Dittrich, J. Krogstie, A.L. Opdahl, S. Brinkkemper, Data integration — problems, approaches, and perspectives, Conceptual Modelling in Information Systems Engineering, Springer, Berlin, Heidelberg, 2007, doi:10.1007/978-3-540-72677-7_3.

[12] N.T. Le, J.W. Wang, D.H. Le, C.C. Wang, T.N. Nguyen, Fingerprint enhancement based on tensor of wavelet subbands for classification, IEEE Access 8 (2020) 6602–6615.

[13] B.S. Saini, C. Khosla, Enhancing performance of deep learning models with different data augmentation techniques: a survey, in: Proceedings of the International Conference on Intelligent Engineering and Management (ICIEM), IEEE, 2020 978-1-7281-4097-1/20/$31.00 ©.

[14] Z. Guo, K. Yu, Y. Li, G. Srivastava, J.C.W. Lin, Deep learning-embedded social internet of things for ambiguity-aware social recommendations, IEEE Trans. Netw. Sci. Eng. (2021).

[15] B. Gorad, S. Kotrappa, Novel dataset generation for Indian Brinjal plant using image data augmentation, in: Proceedings of the IOP Conference Series: Materials Science and Engineering, 1065, 2021, doi:10.1088/1757-899X/1065/1/012041.

[16] H.T. Duong, T.A. Nguyen-Thi, A review: pre-processing techniques and data augmentation for sentiment analysis, Comput. Soc. Netw. 8 (2021) 1, doi:10.1186/s40649-020-00080-x.

[17] P. Subramani, B.D. Parameshachari, Prediction of muscular paralysis disease based on hybrid feature extraction with machine learning technique for COVID-19 and post-COVID-19 patients, Pers. Ubiquitous Comput. (2021) 1–14.

[18] C. Song, Analysis on the impact of data augmentation on target recognition for UAV-based transmission line inspection, Hindawi Complex. (2020) VolumeArticle ID 3107450, 11 pages, doi:10.1155/2020/3107450.

[19] D.L. Vu, T.K. Nguyen, T.V. Nguyen, T.N. Nguyen, F. Massacci, P.H. Phung, HIT4Mal: hybrid image transformation for malware classification, Trans. Emerg. Telecommun. Technol. 31 (11) (2020) e3789.

[20] O'Gara S., and McGuinness K., "Comparing data augmentation strategies for deep image classification", IMVIP 2019: Irish Machine Vision and Image Processing, Technological University Dublin, Dublin, Ireland, August 28-30. doi:10.21427/148b-ar75
[21] J. Zhang, K. Yu, Z. Wen, X. Qi, A.K. Paul, in: 3D Reconstruction for Motion Blurred Images Using Deep Learning-Based Intelligent Systems, 66, CMC-Computers Materials & Continua, 2021, pp. 2087–2104.
[22] B. Gorad, S. Kotrappa, Novel dataset generation for Indian Brinjal Plant using image data augmentation, in: Proceedings of the IOP Conference Series: Materials Science and Engineering, 1065, 2021, doi:10.1088/1757-899X/1065/1/012041.
[23] Kenji Iwana B., "An empirical survey of data augmentation for time series classification with neural networks", 10.1371/journal.pone.0254841 July 15, 2021
[24] Nemade, D. Shah, IoT based water parameter testing in linear topology, in: Proceedings of the 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020, pp. 546–551, doi:10.1109/Confluence47617.2020.90582.
[25] xxx 2022 Figure 2 source - BAB II.pdf (uny.ac.id)