ABSTRACT:

In this research work, we evaluated data augmentation techniques on honeybee classification dataset. We used various image transformation techniques which could be potentially applied for data augmentation. We also used generative adversarial networks GAN for data augmentation and evaluated the performances of generative adversarial networks with image transformation techniques used for data augmentation. Our first proposed methodology of data augmentation accurately identifies distinct, increased images 95% of the total of something and are used in subsequent disease classification methods. This proposed approach is characterised as image processing with both the intention of eliminating infit, unclear, low light X-Rays & producing a data adequate to disease identification. Our second proposed approach is to train a class of GAN’s network on existing honeybee images, in which we used 2 neural networks architectures one is generator, and one is discriminator, after train both architectures against each other. Later, I used the trained generator architecture to generator images for data augmentation. I trained GAN’s with initialize configurations and evaluated the performances of the GAN’s network with image transformations techniques for data augmentation. We also analysed the performances of 2 case study of data augmentation in which our performance varies based on data augmentation techniques we used to improve the performance of CNN in predicting classes for images.

We proposed that data augmentation technique could be improved a model performance in increasing the size of the datasets, which could indirectly reduce the quantity of imbalance in the dataset. When data set is large enough, then it will not significantly improve the model performance, while if datasets are small enough, then it will significantly improve the model performance. Apart from that we proposed that selection of different data augmentation technique could also play a major role in improving model performance. As one of the data augmentation techniques transformations outperformed the best in all the other data augmentation in terms of model performance.

**Introduction**

In this report, we introduced the application of data augmentation techniques on a honeybee classification dataset. At first, we presented some background research about synthetic data and identification using data augmentation techniques. Later we also did background research on various researchers how did they performed data augmentation on image dataset. Then we also presented what is data augmentation, why we need data augmentation and the necessity of data augmentation. Then we elaborated how it can be applied on a structured and unstructured data set. Then we elaborated that what is the importance of data augmentation and then we gave a general view on data augmentation techniques which can be applied on initial data of images including geometric transformations, filters, and mixing images then we also provided some other techniques like adversarial training, Gan, neural style transfer networks and reinforcement learning that how to use techniques could be used to achieve the data augmentation on these datasets. Summary for all background research, followed by the data augmentation techniques we used, contributions to the project and performance analysis along with conclusion and future work is given below.

In this work, we comprehensively worked on data augmentation specifically on honeybee classification, we did some background research about synthetic data and how some of the approaches involving generative adverbial neural networks can be used to produce new data. Or some of the transformation techniques also could be used in data generation then we talked about data creation in which we elaborate GAN describe some of the background literature and research using GANs.

After that we talked about the contribution to this project that why honey bee classification are important then applications of honeybee classification then we describe data augmentation then we explain some of the core reasons which lead to the data augmentation techniques then we briefly describe the data augmentation and its application in various kind of structure and unstructured data set then we describe some data augmentation techniques involving geometric transformations in data transformations, random erasing and mixing images, then some of the techniques which could be used in any data set to transform the augmentation. We talk about adversarial training and a short description about generative adversarial neural networks followed by the neural style transfer neural list style transfer talked about some of the algorithms which can be used in manipulating digital images are mutate by shallow or deep neural networks, these can be used for image transformation.

We also put the limelight on various mobile applications like to be part of Prisma which use the neural style transfer at its very core in these types of methods which inspired from neural style transfer led to the foundation of various networks then we talked about reinforcement learning that how reinforcement learning could be used as an application in the documentation.

Then we described the dataset along with the technical specifications which are needed in implementing this experiment and later we comprehensively talked about that evaluation plan for our experiment we also talked about this plan.

**Background Research**

**Synthetic Data:**

To analyse specific effects of Artificial Intelligence on various fields, one must invest in synthetic data. The process of generating artificial data is resource intensive, complex, and time-consuming. A new study in which artificial data is employed must be approved and funded by the Ethics Committee of a research institute. These expensive procedures are time-consuming and often do not yield informative 6results. In this study, it was found that synthetic data can be produced with a high degree of accuracy and without the need for human involvement. Artificial Data Artificial data can be produced using algorithms. They are computer programs or sets of instructions that allow a computer to create new data without any input from an external source. Artificial Data Synthetic Data The accuracy of these products is influenced by the parameterization used. The process of parameterization entails assigning values to numerical variables such as temperature, pressure, and volume, as well as parameters such as time, time-to-break, and velocity.

Via data augmentation, a process through which machine learning (ML) systems with access to natural language datasets - text or images – generate "new" examples by adding, removing, or changing the content in the original dataset. Data Augmentation can be beneficial in industries like pharmaceuticals and modelling where artificial data can help create incremental value. Data augmentation in healthcare: A study published in the Journal of the American Medical Informatics Association examined the longitudinal trajectory of cardiac risk factors through to mortality. The data set representing this longitudinal cohort contained patients with and without myocardial infarction (MI) defined as a heart attack. Researchers modified the unmodified dataset by adding incomplete records where it was possible to determine whether a patient had an MI. The era of the 12-lead electrocardiograph is conventionally regarded as stretching from 1930 to 1965. In truth, it is possible to find electrocardiographs dated back to 1878 and 1885, with some data sets spanning back as far as 1880.However, the earliest electrocardiogram data set that has been found dates to 1898.

Strategy: AI data that is generated artificially helps set a foundation for new experiences that might not have been created otherwise. For example, there are some retailers using Data Augmented images to refine product selection processes at scale with computers imagining new options or variations that weren’t seen by the humans who originally created the process. Strategy: Artificial Intelligence is a tool to help solve problems, not create them. .AI is a tool, not a replacement for human ability.

Up to this point, we've investigated the idea of using technique to separate and change current photographs to produce new ones. In recent research conducted all around the world, the problem of creating vector images has received extensive attention. Big Photoshopped pictures that we see online are one example of this.

To do it too, specific picture attributes were altered utilizing predetermined settings with each distinct purpose. Within contemporary areas of research, any use of Generative Adversarial Network (GAN), that has been well recognised for producing new pictures that seem to be strikingly like existent pictures, has growing.

**Data Creation:**

Generative adversarial networks (GANs) for pictures as well as transfer learning for text are small set contemporary strategies for intelligent data generation that integrate data sets with increased data. If your desire bike images, then may training GANs on well images of bikes to produce fresh yet accurate images of bikes. According to something like this, you may train language models to produce original statements that incorporate terms or discuss subjects. The pretrained models that are utilised in environmental extracted features are frequently like those algorithms. Expert assessment may assist determine which generated data is realistic since neither situation involves data that is always 100% correct.

The sensitivity of training data could be addressed whether information was produced by people or robotic methods. If your linguistic system is based on online collected information and efficiently collects confidential material, also including address, it might also be exposed to reversing programming, that would reveal the addresses. However, you may make a temporary modelling using the additional data if you can rewrite all the sequences using a learning algorithm and verify that perhaps the fresh sequencing doesn’t appear in the source data. That confidential material in such approach should have been significantly more difficult to uncover through decrypting. Although it is not the focus of this book, data sensitivity is highlighted here as a crucial area in which human-in-the-loop machine learning may be used.

Data augmentation is another way to reduce model overfitting by increasing the amount of training data using only the information in the training data. The field of data augmentation is not new, and in fact various data augmentation techniques have been applied to specific problems. The most important techniques fall into the category of data warping. This is an approach aimed at directly extending the model's input data in the data space. This idea can be traced back to the extensions performed on the MNIST set. A very common and accepted current method for augmenting images is to perform geometric and color augmentation like this: Flip images, crop and translate images, change the color palette of images, and more. All transformations are affine transformations of the original image and take the form: y = W x + b This idea is taken further, using data augmentation techniques on each layer of a to generate a new training pattern A 0.35% error rate was achieved by deep network. Numerical data are enhanced by elastic deformations in addition to the usual affine transformations. Moreover, we know that data augmentation can be applied in many areas beyond simply creating data.

It has proven useful for generalizing computer models to real-world tasks. Generative Adversarial Nets (GANs) are powerful techniques for unattended generation of new images for training. It has also proven to be very effective in many data generation tasks such as generating new paragraphs. With the min-max strategy, one neural network continuously generates better fake samples from the original data distribution to trick the other network. Other networks are then trained to better identify fakes. GANs are used for style transfer. Transfer the image of one shot to another shot (CycleGAN). These generated images can be used, for example, to train a car to drive at night or in the rain, using only data collected on sunny days. Furthermore, by applying transfer learning techniques, GANs were effective even on relatively small datasets. Additionally, it has proven to be very good at extending datasets. Increase the resolution of the input image. Finally, we explore how to train a neural network to augment and classify simultaneously. A similar approach was attempted, but that approach learned different weights for combining existing methods. In our case, we can train a style transfer network to learn how best to generate data augmentation. The goal is not only to reduce overfitting due to augmentation, but also to enrich the data in ways that optimally improve the classifier. These methods do not necessarily produce images like the training set, as techniques such as affine transformations and GANs do. Therefore, it saves manual conversion or correlation between the image generated using methods such as GAN and the original image.

**Contribution to the Project**

**Honeybee classification?**

These bees pertain to the Apies genus as well as the Apidae group. Bee’s species were categorized into four categories, one with its own appearance. Every grouping comprising Asian species has been further subdivided onto two or more organisms. Many bee species have the potential can manage bee heat in the brooding homes, protect the brood and their feed via keeping the foodstuff beneath her brood, or survive solely on pollen & honey.

Computers analyse the pollen they pick up when they visit flowers. These pollens provide clues to certain plants in a certain area. It is necessary for honeybees to keep this data augmented to make sure the most accurate data is being analysed and applied for their use. The honeybees pollinate plants for the flowers to produce fruits and vegetables. This is beneficial to agriculture because this provides an increase in the production of some vegetables such as cucumbers, squash, and melons. Without these benefits, it would be necessary to change other aspects of agriculture like growing different types of flowers or changing the amount of plant fertilization. This is beneficial to agriculture because this provides an increase in the production of some vegetables such as cucumbers, squash, and melons. Without these benefits, it would be necessary to change other aspects of agriculture like growing different types of flowers or changing the amount of plant fertilization. Transpiration is a process where plants lose water through their leaves. This can be

used to measure the water needs of plants. It is a process where plants lose water through their leaves. This can be used to measure the water needs of plants. The definition of transpiration seen in an online dictionary is "the evaporation and subsequent condensation on the leaves and stem of a plant”; this means that transpiration is the loss of water from the air. Transpiration is the process in which water molecules (H2O) are released into the atmosphere through evapotranspiration, and these molecules eventually return to earth as precipitation. Transpiration is an important part of the hydrologic cycle because it is responsible for providing water to plants in a manner that replenishes soil moisture. Transpiration is the process whereby plants release water vapor to the atmosphere through their leaves. The water vapor then becomes clouds which form precipitation.

**Why honeybees are important?**

Honeybees are one of the world's finest breeders, which would be significant for said global food supply. Bee fertilization was thought to be the cause to 33% percent food production. Honeybees for approximately 70 of a 100+ grain animals that feed 90percent of overall of the global total. Pollinators were accountable approximately $1 trillion for agricultural output each year.

Honeybees help pollinate flowers and plants, they produce the honey that we like, they give birth to their young ones. They play an essential role in ecology and the ecosystem. We should not think of these AI writers as a replacement for human copywriters. Instead, they just aid the copywriters by getting rid of writer’s block and generating content ideas at scale. AI writers are not a replacement for human copywriters. Instead, they just aid the copywriters by getting rid of writer's block and generating content ideas at scale. .AI-driven content writers are not a replacement for human copywriters. Instead, they just help the copywriters by getting rid of writer’s block and generating content ideas at scale.

Without bees foraging for food, pollinating plants, and building up their honeycombs, humans could not exist on earth.

In artificial intelligence when we use Data augmentation more difficult to recognize objects with simpler shapes become easier to identify because of larger amount of data. points to compare in order to get more data points, we can increase the size of the database by adding more samples, or we can add noise to the dataset and use a machine learning algorithm like k-nearest neighbour classifier. These two options are only viable if you're working with small datasets. As datasets grow larger, they become computationally inf feasible and no longer viable for analysis.

**Applications of Honeybee Classification:**

1) eDNA analysis of Honey

DNA has been pivotal in the process of classification. To understand the genes and organs of different honeybees, researchers had to concentrate on the genetic composition of their irksome DNA. DNA analysis has been used for various levels of data augmentation e.g., Histones for understanding introns, transcription factors for cell differentiations but with time; this is changing. Now depending on your requirement, you can go with a finite number of enzymes

like Alkaline Proteases, Alpha Amylase and numerous other kinetically related enzymes which are based on evolution-oriented research methodology. Bioinformatics is a computational process that allows researchers to take advantage of the enormous amount of data available from experiments and translate it into meaningful information. This includes analysing and/or predicting the dynamics, pathway, structure, and function of molecules by mining databases related to DNA or proteins. Bioinformatics has helped uncover the different mechanisms behind certain diseases or abnormalities which are only now beginning to be understood. In the healthcare field, bioinformatics helps doctors and researchers analyse data from large amounts of different sources, including DNA sequences and patient reports. Bioinformatics is also being used in agriculture, as scientists want to monitor plant health using satellites or drones. Researchers can use this capability to identify plants that have been infested with insects, plant disease, or pests. The biggest barrier is that the focus is on agriculture and not necessarily bioinformatics. A recent example of how bioinformatics has changed detection techniques in the field of medicine is cancer research. Honeybees play a significant role in maintaining pollination procedures as they traverse around landscapes to search plants that constitute major food sources for honey production. These insects contribute to honey by carrying back nectar to the hive and depositing this in special cells where the honey is produced. The productivity of bee colonies depends on their ability to search out food sources, collect nectar and store it as honey. Honeybees are known for their ability to accurately assess risk when they come across new stimuli. This is done by examining the reactions of other bees in a collective manner. The role of honeybees in pollination and honey production has been well documented, as they contribute significantly to the food source necessary for humans. Honeybees play a significant role in maintaining pollination procedures as they traverse around landscapes to search plants that constitute major food sources for honey production. These insects contribute to honey by carrying back nectar to the hive where it is converted into honey. Honeybees also contribute to the hive's energy and protein needs by foraging for pollen and wax or collecting honey from flowers. The pollination process is significantly influenced by the bee population through the co-evolutionary process of plants producing a variety of floral chemical compounds (flavanols, terpenes, monoter penes, etc.) and bees developing a variety of internal physiological adaptations. The bee's' response to these chemicals determines the success of pollination, thereby influencing the flowering rates of the plants. Bees that cannot forage on their own are supported by honey production and managed by apiaries. Honeybee colonies can be owned or rented from beekeepers, who maintain their colonies' health and often supply bees, wax, and other supplies. Honeybees are the only insects that produce honey without the production of a nectar by-product. And this reduction of the by-product helps in decreasing the possible exhaustions that can affect the surroundings or atmosphere like how the by-products of certain mammals are causing harmful effects.

2) Beyond Pollination

Rather than just changing raw data, recent research suggests that more advanced image enhancement built for machine classification can help increase accuracy when demonstrating visual sentiment. analysis. Online sentiment analysis. Deep learning. Deep learning is an area of machine learning that has achieved breakthrough results in image classification and retrieval. A neural network with multiple layers is trained to provide insights into the data to answer questions such as "How happy is this picture?" or "Who are these people?" In recent years, deep learning

has been applied to a wide range of problems in computer vision and natural language processing, from facial recognition to speech recognition. Deep learning is an area of machine learning that has achieved breakthrough results in image classification and retrieval. A neural network with multiple layers is trained to provide insights into the data to answer questions such as "How happy are people at this point in time, relative to how happy are they five minutes ago?"

Beyond Pollination Honeybees assist with the various connections of sentiment data across various media. This includes images, audios, text, and other aspects... of art. Data from the database is mapped and transferred to a 3D representation of a flower that was designed by artist Sarah Lewis with honeybees in mind. One for One: Rescuing Honeybees from Extinction Honeybees are at risk of extinction because of the pesticides used on crops, which leads to their demise.

3) continuous monitoring of honeybee

The classification monitoring tasks are different from other kinds of labelling tasks. The labels are a binary code because they only need to represent whether the result is good or bad. Also, a monitoring task doesn't undergo data annotation, which focuses on the process of manually identifying existing data in large-scale unlabelled datasets. The parameters that mark the success or failure of monitoring are within some more subtle boundaries. What's more, being at the frontline of charting out new fronts and developing latest technologies challenges inspection agencies to find new and sophisticated methods for classification to accomplish their mission as well as efficiently manage resources available. The classification monitoring tasks are different from other kinds of labelling tasks. The labels are a binary code because they only need to represent whether the result is good or bad. Also, a monitoring task doesn't undergo data annotation, which focuses on the process of manually identifying existing data in large-scale unlabelled datasets. The parameters that mark the success or failure of a monitoring task are not human-readable. It is important to note, however, that the distinction between classification and labelling tasks should not be taken too literally. All tasks involve the categorization of data in some way. Labelling tasks can involve the use of machine learning algorithms like Support Vector Machines to produce labels automatically if they are unlabelled, but classification tasks require the human labelling of data without aid from machine learning algorithms. Classification and Labelling Tasks In general, classification is more about using statistical analysis to predict a response for a particular input, while labelling is more about applying values or classifications to things. For example, labelling could be making an assessment as to whether a patient has a particular disease or not. In classification tasks, the answer is typically provided by a machine learning algorithm. The goal of labelling is to match the input values with one of some fixed number of predefined classes. The goal of classification is to predict which category an input falls into based on the information available in the data set. Accordingly, classification tasks are typically solved by using a machine learning algorithm. The output of the machine learning algorithm is known as the label. This will be one of several fixed values such as "bird", "dog", or "cat".

**Why Data Augmentation?**

In artificial intelligence, when our model or say machine learning model doesn't perform well on test data or doesn't generalise well from training data to test data, it is called overfitting. In this problem our model suffers from high variance. We have number of options to deal with this

problem, first solution could be that we can apply cross validation on the dataset. In this method we can on split our dataset into folds and we can train our model on every fold which indirectly equal to that our model will be trained on each example in the training set. Model performance could be measured by calculating the mean of the performance of model on each fold.

The second solution to prevent overfitting can be by introducing regularization terms or introducing regularization in the cost function. This will directly penalize the cost function and would be able to increase the performance of model on test set.

Another solution could be adding more training data to the model, but it doesn't always work as this may also add more noise to the training data. So, adding more data in the training data will not always work so that's where data augmentation comes for rescue.

**What is Data Augmentation?**

Data augmentation is a technique which is used to expand the size of the dataset by artificially adding that modified version of existing data. Data augmentation could prove highly beneficial to reduce the overfitting and improve the model performance significantly. It can also be used when we have a limited amount of data. One of the best use cases in imbalanced dataset. For example, we are doing classification and we have two categories of images but the number of images of one category is significantly higher than the other images belonging to another category. In this case we must balance the dataset. Data augmentation technique can be used to artificially expand the size of the number of images belonging to the category which has a smaller number of images in the existing database.

In general, Data Augmentation is used most when deep learning model used. This is the reason data augmentation techniques introduced with several deep learning frameworks. Apart from this, data augmentation could potentially be used for machine learning problems as well as,

Data augmentation can be used for both structured data as well as unstructured data. The following can be augmented easily:

1. Audio data

Audio data augmentation is one of the most feasible and performant ways to increase the size of training set of speech data. Hence, enabling better system coverage through dataset variations. is one of the most important areas that need attention from the speech recognition field. The term “quality” in the field of speech recognition needs to be understood as having different meanings. A quality refers to some property or aspects of a speech signal, such as its perceived loudness and richness. Quality can also refer to the degree of correctness in transcription and recognition. The concept is similar but not identical with accuracy in that an accurate transcription may not necessarily be of high quality, whereas a poor-quality transcription can be accurate. A quality measure captures an aspect of accuracy and the confidence or trustworthiness of the human-generated recognition results. The concept is similar but not identical with accuracy in that an accurate transcription may not necessarily be of high quality, whereas a poor-quality transcription can be accurate. The term "accurate" is often used in a strict way to mean an exact representation of the original text, or the most similar, but not necessarily the best. That is, accuracy often means

that something is so detailed that it may be taking up space and time without being effective. There are many different definitions of what constitutes accurate transcription and some of them are explained below. A well-transcribed text is one that meets the following criteria: It includes all or most of the words in the original text. It includes punctuation and spacing that is closest to what is used in the original text. It includes sentence breaks, paragraph breaks, and other changes between sections of a text that make sense. A word-for-word transcription is one that includes the exact words in the original text. A well-transcribed text would include all or most of the words in a text but would not be a word-for-word transcription. For example, "He had two pencils" would be considered.

2. Text data

Text data for data augmentation is extracted from news, blogs, and websites with the content has a large quantity of text. AI writer is still new to the industry and have noticeable limitations. in terms of vocabulary. AI writer is still new to the industry and have noticeable limitations in terms of vocabulary. AI writer is new to the industry and has observed limitations in terms of sentence structure and length.

3. Images data

Image data augmentation is the process of modifying existing imagery. Image data augmentation is a key aspect of deep learning training and can be an imperative step to correctly train AI systems on new assignments. Advanced pedagogy research shows that these modifications do not harm AI and do increase the accuracy of image recognition systems. Image data augmentation can be used to increase the size of a network's input image database, add content such as key points, or to take an existing image and improve its quality. It is a crucial step in training deep learning systems. Pix2pix is an example of image data augmentation technology, which has been used to help robots recognize objects in real-time. Dense Net is a deep neural network architecture for image classification that has been improved by photo-Realistic Image Synthesis. In computer vision, image data augmentation is the process of changing the representation of an input image to present it as larger and sharper than it is. It can be used to increase the size of a network's input image database, add content such as textures, generate new images or to smooth images. In computer graphics, image data augmentation is the process of changing a texture pattern to be larger and clearer than it is. It can be used to increase the size of a texture map database, generate new textures, or increase image quality. In both computer vision and graphics, data augmentation often involves sampling an image at different wavelengths and then modifying the pixel intensities to creatively change the image. In computer vision, data augmentation is used to create more realistic images by creating new pixel intensities that don't exist in the original image. In computer graphics, data augmentation is often used to generate new textures or modify old ones. Data augmentation can be implemented using a procedural texture generator or a data-driven technique. Data augmentation is commonly used to generate new textures from an existing one, e.g., rotating and translating a given texture while performing various other operations on it.

4. Any other types of data

In this, we applied data augmentation on images dataset (honeybee images dataset). Everything related to this augmentation is clearly explained in the subsequent paragraphs moving forward.

**Importance of Data Augmentation**

Data Augmentation is useful in improving machine learning and deep learning model performances and outcomes.

The data augmentation tools make the data rich and sufficient and thus makes the model perform better and accurately. Data augmentation techniques reduce the operational costs by introducing transformation in the datasets.

Data augmentation assists in data cleaning, which is essential for high accuracy models. Data augmentation makes machine learning more robust by creating variations in the model.

**Adversarial training**

In adversarial training, also called adversarial machine learning training, they generate adversarial examples by disrupting machine learning models and after disrupting machine learning models, they inject these into datasets.

Adversarial training is a branch of reinforcement learning in which an AI agent tries to see matches what it expects to see. in a game. This can be as simple as a player seeking out and destroying

enemy units, to more complicated scenarios like making sure that it is never in the same place twice during a SAVE game. There are many implementations of this training method such as MT-Q learning, RLSAVE and ISSAVE. The recurrence relation used by these algorithms can be expressed in a matrix form. Adversarial training is a branch of reinforcement learning in which an AI agent tries to see matches what it expects to see.in a game. This can be as simple as a player seeking out and destroying enemy units, to more complicated scenarios like making sure that it is never in the same place twice in a level. The term "adversarial training" was coined by Rolf Pfeifer and was popularized by video game designer Chris Hecker in his book AI: A Modern Approach. Game designers can use adversarial reward functions to create gameplay or as part of story missions.

This approach to AI is used in tasks where the AI has access to some observations that include false observations and positive observations. The false observations are not as important, but once pitted against the positive data, can be used to weaken, or defeat the data making process.

**GANs**

GAN network also known as a generative adversarial network, in this network 2 neural network architectures one is generator and other is discriminator train against each other and after training both networks against each other, the generator is used to generate the new examples and later for data augmentation, we can merge the new examples with the existing database.

Rigorous data grounding aids in training the architectural components of rich generative models.

Generative modelling is shaping up to be an important aspect of generative adversarial networks to speed up the creation and application of models in machine learning. Initial implementations of generative modelling in GANs have focused on generating data from complex distributions such as a categorical distribution, but the real power of generative modelling is that it can be applied to any type of distribution. For example, models for images could be generated for the range [0.5, 1] pixels with a Gaussian distribution (e .g. the probability of a pixel being in the centre is 0.5 and is scaled by 1-x to get x) or for a range [0.7, 0.8] pixels with a square distribution (e.g. the probability of a pixel being in the centre is 0.75).As with other types of generative models, , generative modelling can be combined with other types of models to produce more complex output. For example, the probability of a pixel being in the centre could be a mixture of an exponential and Gaussian distribution. A generative model can also have multiple outputs that depend on certain parameters. For example, a car image could have various features such as wheels, engine compartment, and doors. In general, the outputs from a generative model are discrete categories of information. The input parameters to the model are continuous variables that go into the probability distribution for generating each category of output at a given instant in time. This distribution can be estimated from data or computed using discrete modelling techniques such as Markov chains or Hidden Markov models. For continuous variables, the probability distribution of an outcome given a particular input is called the prior distribution.

**Neural style transfer**

Neural Style Transfer refers to a class of software algorithms that manipulate digital images or videos to obtain the look or visual style of another image. The NST algorithm is characterized by using deep neural networks for image transformation. A common use of NST is to create artificial works of art from photographs, such as converting the look of a famous painting into a user-provided photograph. Several notable mobile apps such as DeepArt and Prisma use NST technology for this purpose. This method has been used by artists and designers around the world to develop new artwork based on existing styles

**Reinforcement learning**

Reinforcement learning is the training of machine learning models to make a set of decisions. Agents learn to achieve goals in uncertain and potentially complex environments. In reinforcement learning, artificial intelligence faces a game-like situation. Computers find solutions to problems through trial and error. To make machines do what programmers want, artificial intelligence receives either rewards or punishments for actions taken. His goal is to maximize the total bonus. The designer sets the reward policy (rules of the game) but does not give the model hints or suggestions on how to solve the game. From completely random attempts to sophisticated tactics and superhuman abilities, it's up to the model to figure out how to perform tasks that maximize rewards. Leveraging the power of search and lots of experimentation, reinforcement learning is currently the most effective way to hint at the creativity of Hint His Engine. Unlike humans, artificial intelligence can glean experience from thousands of parallel game runs if the reinforcement learning algorithms are running on a sufficiently powerful computing infrastructure.

Reinforcement learning refers to a class of machine learning methods that was designed and developed with the idea that people or animals learn from the experience. of making decisions. This form of learning is driven by the desire to maximize rewards and minimize punishments over time. Neural networks are one of the most popular reinforcement learning techniques in use today. Reinforcement learning is a type of machine learning that helps make decisions based on consequences. The AI learns how to complete tasks such as playing games by trial and error. Evolutionary algorithms are also a popular technique used in reinforcement learning. The process of training an AI or ML algorithm is often referred to as reinforcement learning. Machine learning techniques such as reinforcement learning are important in artificial intelligence and cognitive systems because they allow algorithms to learn from a large amount of data, allowing them to make predictions that people cannot.

There are two main types: Q-learning and SARSA. Q-Learning is when an agent explores its environment while SARSA is where an agent predicts how rewarding a possible action would be, out which it executes it once to study the effect of its behaviour on its environment. The other algorithms are: -DQN: Deep Q-learning which is a variant of Q-learning which improves on the performance of decision making by learning more about the environment. -Policy gradient methods: This is where an agent chooses a policy that maximizes some objective function, where the optimal value of this objective function changes with time. The agent then uses backpropagation to find the gradient of the objective function. It's a supervised algorithm, meaning that it requires training data to learn what is good for the agent. -Reward-switch: Reward-switch algorithm is when an agent learns "if I do A, then I get a reward R. If I do B, then I get a reward Q".

**Requirements**

Data Image pre-processing:

 Image Resizing, Image Rescaling, Image Centre Crop

Image Data augmentation:

 Random Crop, Flip, Translation, Rotation, Zoom In, Zoom Out, Height, Width

 Brightness

 Sharpness

 Contrast

 Tilt an image by another extent.

 Mixing Images

A dataset for data augmentation is a type of dataset that is constructed with new samples or observations in addition to the data from which the dataset was originally generated.

Data Augmentation: A body of datasets which employ data augmentation, may also be referred to as 'data ensembles.' Data augmentation is used when more data has been added to an existing set. For example, if we are attempting to train a classification algorithm using some small set of pixels and we could add more pixels on either side, then the resulting larger set will represent a full search grid around each pixel in the original set with similar process iterations preventing overfitting converging faster.

**Dataset Description**

https://github.com/jaddoescad/ants-bees-dataset/tree/master/train.

**Technical Specification**

To make the neural network effectively and consistently produce the desired result, we need to provide it with as much training data as possible. Data augmentation uses generators for new examples with augmentations which are permutations, translations, reflections, and other transformations of input examples. Augmentations add new pixels that weren't present in the original data and can't be seen through resampling.

Let's look at how to use keras for data augmentation. The procedure that we must employ is

ImageDataGenerator.

Here are just an almost all the various augmentation methods offered by Keras APIs that is being used to generate extra data for network training.

Initially, you must import it from keras\_preprocessing. image:

Crop\_image = tf.image.central\_crop(.jpg , central\_fraction=0.7) > For Cropping Image

Crop\_and\_resize\_image = tf.image.crop\_and\_resize(.jpg) > For Cropping and Resizeig the image.

Flipping\_image = tf.image.flip\_left\_right(.jpg) > For Flipping Image

Rotate\_image = tf.contrib.image.rotate(.jpg ,math.radians(270)) > For Rotate Image

**Requirements Evaluation Plan**

The goal of Data Augmentation testing is to verify the quality of results generated. Therefore, data scientists are suggested to strive for accuracy, functionality, security, and usability to verify the software's quality.

Data augmentation testing commonly focuses on three parameters: Static Data Verification, Dynamic Data Verification and Evaluating Content Quality. Testing should be done under logs containing a mix of stratified data with one or more audio files or video files.

There are different ways a person can verify if a project has been successful. I will compare my results with past lessons such as field incidents where people have had success at augmentation and leading studies in data augmentation/anticipating needs and habits of consumers/competitors.

Standardizing your information is crucial to ensuring that the Data Augmentation APIs and methodologies you really want apply result throughout the desired results. The necessary phases could be used to properly verify your picture data set.

 Check to see if each of the augmentation APIs is generating extra photos.

 Check to see if the photographs produced fall together under the categorization.

 Check to see whether you might not have the same photos repeated by observing the

 changes every API performs towards the pictures.

**Risk Plan**

Data augmentation is not a new concept, but its importance becomes more and more sophisticated with the development of machine learning (ML) algorithms.

The ML algorithms can learn to identify which areas of data might be missing or anomalous and use math to find ways to fill those gaps. What are the necessary steps for data augmentation?

That will make deep learning (DL) training process more robust due to high diversity within training dataset as well as improvement of classification accuracy as ML can identify anomalies which could be unrecognizable otherwise.

The risk plan is the first thing professionals will implement if they intend to take their data marketing and management to a profession and reliable level.

The risk of this data augmentation is understanding the image clearly by using ML techniques.

By using the techniques, we should know that how the Honeybee is and where is it is lying.

**Methodology**

**Initial Data Augmentation**

The Augmented API is one such tool that helps in scaling data augmentation. It allows users to add missing or inaccurate data to their datasets, as well as make changes to existing data without losing any of the existing information.

**AddNoise** Adding noise to an image is useful when the image is blurry. Salt and pepper noise makes the image look like it is made up of black and white dots.

Data augmentation is an analytics technique that helps in presenting the data in simpler and easier to interpret visual form. It helps in presenting meaningful information with the accuracy attained by using less amount of data.

Add Noise primarily uses testing methods to filter out rare items within a subset of user recall values on success. The results are shown in the form of distributions.

In Data Augmentation, this technique helps us filter for some items in our dataset and show output based on these filters.

**Crop** Select a portion of the image to crop and scale to the original image size.

Data Augmentation is a technique that allows the machine learning algorithms to learn from the data it has been fed. This technique is used in crop techniques.

Crop techniques are a type of machine learning algorithm, which uses data augmentation to improve the accuracy of its predictions.

Crop methods are primarily used for image classification and segmentation problems, in which images must be divided into different categories based on their content.

**Flip** Flips the image horizontally. Mirroring rearranges pixels while preserving image features. Flipping upside down doesn't help in some pictures, but it does in cosmology and microscopy.

Data augmentation is a technique that helps in improving the quality of data. It is also known as Data Flipping. In this technique, the data that has been generated by some source is flipped and then used to generate new data.

Data Augmentation can be applied when you are using different sources of data for your model and want to get an idea about how different sources of data affect the model's performance. **Rotate** Image between 0 and 360 degrees he rotates 1 degree. Each rotated image is unique within the model.

Data Augmentation is a technique that can be used to make an existing dataset more robust. It helps in improving the quality of data by adding new data points and increasing the number of features.

The Rotate technique is a Data Augmentation tool that works on the principle of "rotating" different classes or features to generate more data points. This technique is applied when there are too few data points and too many features for the desired outcome.

**Scaling** The image is scaled outward and inward. Objects in the new image may be larger or smaller than the original image due to scaling.

**Translate** The image is translated along the x or y axis to different regions and the neural network searches and captures everywhere in the image.

Brightness

Changes the brightness of the image, making the new image darker and lighter. This technique allows the model to perceive images at different levels of illumination.

**Contrast** The contrast of the image is changed, and the new image has different brightness and color. Randomly change the contrast of the image below.

**Advantages and Disadvantages of Data Augmentation**

**Advantages of Data Augmentation** Reduces data collection costs. Reduce data labeling costs. Improves the predictive accuracy of the model. Prevent data shortage. Design a better data model. Reduces overfitting of data. Brings variability and flexibility to your data model. Improves the ability to generalize the data model. Helps solve class imbalance problems in classification.

**Limitations of Data Augmentation** Let's discuss the limitations and challenges of data augmentation along with its benefits. Data expansion limits are as follows: We need a system to assess the quality of the augmented dataset. New research is required to create new or synthetic data using advanced applications. It is very difficult to apply some data augmentation techniques such as GAN. Identifying the best data augmentation strategy is another challenge. Augmented data contains a bias when the actual data set contains the same data. Data Augmentation Applications In the medical field, data augmentation is used in the field of medical imaging. In the medical field, it is used to discover rare diseases. Example:

1. to identify brain tumors.

2. For differential data expansion.

3. For automatic data enrichment.

4. For partially supervised work.

In the technical field, data augmentation has the following applications: Deep learning software. Natural Language Processing. Image Recognition Machine Learning Software.

**Performance Analysis**

**Case Study 1**

We analyzed a case study [C1] on performance analysis of model with data augmentation and model without data augmentation. This case study used Fashion MNIST data which has 60, 000 training images and 10,000 test images. On this dataset, they built a convolutional neural network model which is used on both data with augmentation and data without augmentation. After training the model on training data without augmentation and testing on test set. They measured 4 performance matrices training loss, validation loss, training accuracy validation accuracy. After applying data augmentation, they used the same convolutional neural networks architecture.

**Case Study 2**

In this case study, we analyzed a CNN model by testing its performance on data after applying different data augmentation techniques and the same data without applying the data augmentation techniques. In this experiment the data which has been used is flowers dataset with 20 categories. CNN model which was used in this is Resnet 18 model. Resnet 18 model has been trained on ImageNet. So technically in this study they used three data augmentation techniques and measured the performance on data without argumentation. This means that we have a model performance on four different types of data set, three of which are augmented on the same data set, and one is the original dataset.

Resnet 18 CNN model is a pretrained network which was trained on the Image net dataset. During training the data set was normalized separately for each color, so we normalize the mean and standard of the images according to the network on which the data set is trained on. So, the objective of normalization is that we must ensure that color channel is zero centered with a range from -1 to 1.

And then it is followed by a label mapping to categorize the label with the category of the flower so they use Json object which can be read using the Json mapping so it’s basically using a dictionary mapping to map to the categories to the actual name of the flowers. As in this experiment we have 20 categories of flower, so we need to customize the ResNet 18 so by applying the SoftMax function with a linear output of 20 with customize the ResNet classifier with linear output of 20.

So, they applied three types of data augmentation on the dataset one by one separately. First, they applied first augmentation where they augmented the data using a random horizontal flip to transform the training set only and after augmented the data using random horizontal flip in training data by training our resident 18 model on this first augmented data and then measured the performance on the test set.

The second argumentation technique they used is introducing a random rotation by a factor of p equal to 0.5. So how can we achieve this augmentation? This can be achieved by transforming the current transform object with the interaction of random rotation with a multiple of 0.5. Apply the same process of training by Resnet on this type second augmented data technique followed by measuring the performance of the Resnet on the test set.

The third data augmentation techniques were used is gladiator augmentation technique so what we can do is we took the training data and transform the training data using **ColorJitter** augmentation technique. After transforming the training data using third argumentation data technique using ColorJitter, the Resnet model has been trained on this third augmented data and measured the performance on a test set.

After measuring the performance of Resnet 18 model on the data set without augmentation and the three data set which are derived from the data argumentation techniques which was named Type 1, Type 2 and Type 3, so we analyzed validation accuracy of Resnet models on all the flowers dataset, one base dataset and three derived datasets from the respected data augmentation techniques. We concluded that the data without augmentation performs the minimum and the validation accuracy across the epochs from 0 to 30 by Resnet model 18 on data without documentation are always below the model performance on the rest of the data derived by all the three data augmentation techniques.

To evaluate the performance of Resnet 18 model on test set on different data configuration. We analyzed that Resnet model on data without augmentation which achieved an accuracy below 65% on test set. While the Resnet 18 model trained on the first data augmentation technique achieved an accuracy on test set close to 70% which can be considerably a significant improvement. Apart from this there is a significant improvement in accuracy on test set by Resnet 18 when we applied the secondary data argumentation techniques or type second data augmentation technique on the test set. Surprisingly, type second data argumentation techniques data trained Resnet model achieved a considerable accuracy of nearly 75% which is better than the first two configuration data with augmentation and the data without data augmentation. Type III augmented data or data obtained by applying the type third augmentation technique on original data, Resnet 18 model trained on this data outperformed the best in all the four models and achieved an accuracy nearly 80% which was a considerable improvement than the type second augmentation and showed a huge jump in the model performance with compared to the data without augmentation.

**Results Interpretation from Case Study 1 and Case Study 2**

After analyzing the case study, first we concluded that performance metrics couldn’t significantly improve as we are expecting after applying the data augmentation technique, we concluded that this is because the model we trained in case study 1 has been trained on the data set more than 7 million images which means that we already have enough data to train any CNN model after applying any data argumentation technique on this big data. There was only a slight improvement in the model's performance, just considerably improved by 2% on the test set.

After analyzing the case study second, we concluded that after applying the different kind of data augmentation techniques there is always room off significant improvement in the model performance. In that case study we have a flower dataset of 20 categories, and we choose a pretrained Resnet model and customized the Resnet 18 CNN model according to the 20 linear outputs. Which means, we need an output layer of SoftMax. After applying 3 data augmentation techniques name type 1, 2 and 3 and apply this Resnet model or train this Resnet model on the, after trained model on 3 data configuration and we found out that yes there is a huge improvement when it comes to validation accuracy or test set. In fact, various data augmentation techniques have also major differences in achieving accuracy on test set. What we concluded from both case studies is that if they have a smaller number of images in a data set or say we have a small data set then there is a high chance of significant improvement in the model performance on data set with compared to larger data set. So, we can see that if we have a small dataset, data augmentation technique can be proven highly effective and after applying the data augmentation techniques only small data set, we can achieve a significant improve the accuracy on validation test set which can be showed in the above two case studies.

Next question arises is why all the data augmentation technique have different accuracies as we can see in cases study 2, type 3 of data augmentation techniques achieved the highest considerable accuracy on augmented data, the question is should we still care about the data argument techniques that which data augmentation technique can be proved best when it comes to predicting classes on a test set.

**Selecting the Best Augmentation Techniques**

As we can see in the case study involved different types of data augmentation, we concluded that different data augmentation techniques contribute differently when it comes to achieving considerable accuracy on test set. When it comes to transforming the data set using random horizontal flip, it achieved better accuracy than base data. When the data augmentation technique of random rotation with some factor by using current transformation object, it achieved a different accuracy but more than 1st. Later we used type third data augmentation technique which basically use the color jitter augmentation techniques or applied color jitter augmentation technique to transform the data, it achieved the highest accuracy and outperformed best in all the data configurations. We concluded that we could create our choice of data argumentation technique to come up with the best model trained on the augmented data to select best predictive model.

So, a performance evaluation of the same model architecture on the base data and the augmented data using different augmentation techniques will be a great choice when it comes to selecting the

best data augmentation techniques. This will not only help in improving the model accuracy based on the data said but it will also select the best of the augmentation techniques we have to our hand.

**GAN performance analysis on Honeybee dataset**

So, in the last analysis, we used generative adversarial neural network for the data augmentation. For building GAN, we use TensorFlow. At first, we mount the drive as we are importing the data from Google Drive, then we initialize the configuration. At first, we initialized, the resolution factor equals to 3 which means when 1 is equal to 32, two is equal to 64, three equals to 96, four is equal to 128, so it basically means generative resolution factor is a multiple of 32 then we define generative square which is equal to 32 into generating resolution factor with three channels. We initialize the seed size to generate images which is equal to 100 then we defined the data path apart from which we are going to extract the data and followed by batch size of 32 with a buffer size of 60,000. Then we defined the epochs according to our computational capacity. Then at very next our goal was to reduce the computational time for loading images when often required for this. We just directly generate the squares and convert this into NumPy binary data and then save locally at the same path where all the images were saved. So basically, we named this data as training data. Then we uploaded this data into the train data set using a buffer size of 60,000 with a batch size of 32.

After converting the image data into a portable data we started building two sequential model one generator and one discriminator, so what we will generate a new generator to generator data and discriminator to discriminate the original data from the generated data so objective of discriminator and generator is in that process we have to generate the data in such a manner that we have to fool discriminated so discriminator will not be able to discriminate on that level that it can discriminate fake data and real data. So, we defined a python function for generator other one for discriminator. The generator will take seed size and channels as input while discriminator will take input size of image as input.

After training the GAN network, we can generate the images using that generator function by inputting the random noise. We can improve the images generated by GAN networks by increasing the number of epochs on which both networks trained.

**Conclusions and Future Work**

In this, we described almost all available data augmentation techniques to transform honeybee datasets, we described flipping, random horizontal flip and random vertical flip, image resizing, random zoom, random rescaling, adjust contrast and adjust brightness. Then analyzed 2 case studies which concluded that how data augmentation proved more effective when applied to small datasets as compared to larger datasets. GAN’s came with one of best strategies when generating images based on generator and discriminator models.

**References**

[1] https://www.tensorflow.org/

[2] https://www.tensorflow.org/tutorials/images/data\_augmentation

[3] https://link.springer.com/chapter/10.1007/978-3-658-14577-4\_3

[4] https://ieeexplore.ieee.org/document/8388338/authors#authors

[5] https://towardsdatascience.com/top-python-libraries-for-image-augmentation-in-computer-vision-2566bed0533e

[6] https://neptune.ai/blog/data-augmentation-in-python

[7]https://analyticsindiamag.com/image-data-augmentation-impacts-performance-of-image-classification-with-codes/

[8] C. Chadebec, E. Thibeau-Sutre, N. Burgos and S. Allassonnière, "Data Augmentation in High Dimensional Low Sample Size Setting Using a Geometry-Based Variational Autoencoder," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, doi: 10.1109/TPAMI.2022.3185773.

[9] J. Shijie, W. Ping, J. Peiyi and H. Siping, "Research on data augmentation for image classification based on convolution neural networks," 2017 Chinese Automation Congress (CAC), 2017, pp. 4165-4170, doi: 10.1109/CAC.2017.8243510.

[10] T. Huynh-The and D. -S. Kim, "Data Augmentation For CNN-Based 3D Action Recognition on Small-Scale Datasets," 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), 2019, pp. 239-244, doi: 10.1109/INDIN41052.2019.8972313.

[11] S. Perche, D. Botina, Y. Benezeth, K. Nakamura, R. Gomez and J. Miteran, "Data-Augmentation for Deep Learning Based Remote Photoplethysmography Methods," 2021 International Conference on e-Health and Bioengineering (EHB), 2021, pp. 1-4, doi: 10.1109/EHB52898.2021.9657650.

[12] K. G. V. S. Karthik, K. Nithin, B. Dhanush, K. Praveen and S. Sarath, "Data Augmentation of Neonatal Thermal Images Using Deep Learning," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579769.

[13] https://www.analyticsvidhya.com/blog/2021/04/generate-your-own-dataset-using-gan/

[14] https://www.tensorflow.org/tutorials/images/data\_augmentation