Leveraging Data Science in Cyber Physical Systems to Overcome Covid-19

Harshil Jhaveri[1], Himanshu Ashar[2] and Dr. Ramchandra Mangrulkar[3]

[1,2]Dwarkadas J. Sanghvi College of Engineering, Mumbai
[3]Associate Professor, Dwarkadas J. Sanghvi College of Engineering, Mumbai

Abstract: *As of July 2020, the total cases for the Novel Corona virus, Covid-19, peaked at a massive 12.3 Million victims, with over 550,000 casualties worldwide. In response to the staggering death rate and contagiousness of the disease, several disciplines of Cyber Physical Systems have provided heuristic solutions to flatten the curve and limit the rising cases per day. Big data analytics will act as a medium for tracking, controlling, research and prevention of COVID-19 as a pandemic. COVID-19 can be detected via information compiled using a framework for mobile phones. For Forecasting the time-series data, various DL methods are used to train data in the structured as well as unstructured format, with a biological information systems approach, to create knowledge platforms for research professionals. In order to carry out Detection of COVID-19, pre-trained models as well as customized Convolutional Neural Networks (CNNs) are trained using open-source CXR and CT scan image datasets, to compute their features. COVID-Net is a recent, publicly available, CNN-based model, used to detect COVID-19 in individuals, trained on a dataset of chest X-ray (CXR) images. Social Media can serve as a major source of relevant information on a daily basis. Researchers have conducted multiple studies related to social media analysis, tweets related to COVID-19 owing to the disease's widespread nature.*

Keywords:- ***CNN, Covid NET, Cyber Physical System, Covid-19 Detection, Deep Learning, Big Data.***

1. INTRODUCTION

The novel corona virus, known as the severe acute respiratory syndrome corona virus 2 (SARS-CoV-2), whose first case was observed in December 2019, has had a significant impact on the global population. As of August 4, 2020, there are 18,142,718 confirmed cases of COVID-19, with 691,013 reported deaths across the globe. Symptoms of varying intensity have been reported, ranging from fever and dry cough to less common occurrences such as skin rash and diarrhea. The fatality rate has been reported highest among those suffering from preexisting conditions such as cardiovascular diseases and diabetes. There has also been a significant impact on people's mental health, with health-care providers and medical professionals suffering from stress, anxiety, and other depressive symptoms, along with the general population showing an observable increase in major depression. The pandemic continues to rage around the world, with a large number of countries still reporting a steady increase in the number of cases on a daily basis.

In order to curb the effects and control the spread of COVID-19, there is a need for effective and impactful measures to be implemented. Using a data-driven approach to leverage various algorithms can prove to be very resourceful, provided the data collected is used responsibly, by maintaining user data privacy. Data Science can play a pivotal role in the detection,

control, as well as prevention of corona virus. While various studies are being carried out to develop potential cures for COVID-19, such as vaccines and various forms of treatment, Artificial Intelligence shows a measurable contribution in bringing the situation under control. Screening of patients, predicting the future trend of the virus with the help of time-series data, diagnosis of COVID-19 using medical images, interpreting the public's response to the various changes in daily routine, and real-time spread tracking have been some of the applications of Artificial Intelligence related to COVID-19.

Broadly, Data Science has been used to predict the impact of this disease. It can also be used to conduct simulations related to the spread of COVID-19, appropriate hospital protocols, and effective strategies for observing social distancing. Social Media Analysis has been conducted, with emotional and sentiment analysis conducted based on data scraped from Twitter, related to the public's opinion regarding the lockdown and easing its restrictions, along with detecting communities. Big Data has made significant contributions during this pandemic, such as collecting suspected patients' travel histories, making records of its symptoms and handling information related to COVID-19. Big Data working with Artificial Intelligence can also enable the study of antibody sequences, which could play a role in finding an effective cure. Mobile phone-based surveys can also be conducted to understand the distribution of COVID-19 across various cities and towns in quarantine. Computed Tomography (CT scans) and Chest X-rays (CXRs) are essential for creating Convolutional Neural Networks in order to classify patients as COVID-19 positive or not. Since COVID-19 is a recent situation, the amount of data available from CXR-images may tend to fall short of ideal requirements. In order to generate augmented and effective data, Auxiliary Classifier GAN is used. Deep Learning plays a significant role in the process of detecting positive cases of COVID-19. There are various Neural Network architectures, based on deep CNNs, GANs, LSTMs and others, which are used to detect COVID-19 from medical images. CT scan images and CXR images are the two types of data used for training the deep learning models. Artificial Intelligence also involves machine learning based approaches used for forecasting the spread of COVID-19.

In this chapter, various methods are explored in the field of Artificial Intelligence and Data Science used in order to track and predict the potential spread of COVID-19 through various means, thus helping to prevent further spread of the pandemic. The role of social media analysis techniques which can be beneficial to gather data related to the public and observe their reactions to the changed situation that has been brought about due to the lockdown is studied. A detailed comparison is made of various deep CNNs and Auxiliary GANs used for data augmentation as well as detecting of COVID-19 among the public. We then propose an effective methodology, which can help to successfully detect cases using various neural network architectures, along with data in the form of Chest X-rays or CT scan images. It also includes solutions to extend the dataset by creating synthetic data, with the help of Generator Adversarial Networks (GANs).

The chapter is organised as Section 1 giving detailed information about Covid-19 Infection and spread, data available about symptoms, victims and curve for spread of the infection. Section 2 elucidates related literature found for the usage of Big Data methods for mitigating Covid-19, deep learning models trained on rising trends to forecast the spread of infection, use of CNNs trained on relevant medical images for effective diagnosis of Covid-19. Section 3 explores proposed methodologies of using Mobile Phones Cloud for Big Data mitigation of Covid risks, Covid-Net Neural Network to forecast and predict the rise in cases, Image Nets to diagnose a potential patient, Social Media Analytics for analysing real-time public sentiments. Section 4 gives the proposed implementation of the model. Section 5 later gives

the results and model accuracy obtained after the experiments. The chapter then ends with the conclusion of these methods and on a note for future extensive research.

2. Literature Review

Linda Wang et.al. [1] proposed a Deep Convolutional Neural Network for detecting COVID-19 from chest x-ray (CXR) images. Computed Tomography (CT) scans, and more commonly CXRs have emerged as an effective screening method for detection of COVID-19. This is because CXRs are a standard equipment, which can be easily accessed. Compared to fixed CT scanners, the conveniently portable CXR systems enable testing patients in isolated areas. CXR imaging can be performed parallelly with viral testing, to control the large number of patients. Such diagnostic systems relying on computers are resourceful to assist radiologists in examining CXR and other medical images. Their study introduced an open-sourced deep Convolutional Neural Network, designed for detecting cases of COVID-19 from CXR images, called COVID-Net. They also developed an open access dataset called COVID-x, which included data from 13,975 CXR images related to 13,870 patients. Their prototype for the design of the neural network involved making three predictions: absence of any infections, bacterial and viral non-COVID-19 infections, and COVID-19 infections. This three-way prediction system enables medical professionals to decide the strategy for treating patients based on their type of infection, and prioritizing patients for PCR testing, to procure further information in potential COVID-19 cases.

Abdul Waheed et.al. [2] proposed a system which could generate data from chest x-ray (CXR) images, augment it and detect COVID-19 in patients. It is difficult to obtain sufficient data of CXR images, as COVID-19 is a recent occurrence. The dataset to be trained is artificially extended by using data augmentation methods. GAN is an innovative model, which performs more augmentation than brightness improvements, scaling, flipping etc, to generate synthetic augmented data, without being supervised. GANs use two opposing networks, G(z), a generator which produces realistic images with the purpose of tricking D(x), the discriminator, using which true and false images are distinguished. GANs use a min-max game, as G(z) aims to minimize V(D,G), the cost function, while D(x) is designed to maximize it. Along with an Auxiliary Classifier Generative Adversarial Network (ACGAN), their research proposes a Convolutional Neural Network architecture for the detection of COVID-19. CovidGAN further improves the detection of COVID-19, by augmenting the training set images.

Zhongyi Han et.al. [3] proposed an attention-based deep 3D multi-instance learning (AD3D-MIL) system for determining COVID-19 cases. In a worldwide attempt to curb this pandemic, a lot of research has been conducted in AI powered technologies related to analyzing medical images. A notable method for this purpose is automated screening, which is an important utilization of Artificial Intelligence in COVID-19. Their proposed system has weak labels, but higher interpretability. AD3D-MIL combines three functions, which enhance the interpretability and generalization of screening algorithms; a deep instance generator for automatic generation of deep three dimensional instances, an attention-based MIL pooling for creating an informative bag representation by combining deep instances, and a transformation function, which creates a Bernoulli distribution by transforming the bag representation. Results indicate a higher yet interpretable performance by AD3D-MIL on CT examples. A

thorough analysis has shown that AD3D-MIL is an effective clinical tool, which serves the purpose of abating the amount of work for radiologists in screening COVID-19 cases and streamlining the process.

Md Zahangir Alom et.al. [4] proposed a multi-purpose system for detecting COVID-19 and localizing the infected region by using two medical imaging modalities, involving an end-to-end approach. Their improved versions of the Inception Recurrent Residual Neural Network (IRRCNN) for detection of COVID-19, and NABLA-3 Network models for locating infected segments, by using open-access datasets containing CXRs, abdominal as well as full body CT images. The experimental results displayed an accuracy of 84.67% on CXR and 98.78% on CT images for the test set. For both CXR as well as CT images, high accuracy was obtained to segment and detect regions affected by COVID-19. In order to create a more reliable system, the authors intend to collect more data from COVID-19 affected patients.

Arni S. R. Srinivasa Rao et.al. [5] proposed a machine learning model to help identify COVID-19 cases using a mobile phone-based survey. As indicated by several studies, Artificial Intelligence and Deep Learning can have useful applications to assist the diagnosis and decision-making process to treat various diseases. Website-based portals, along with mobiles, have been successfully tested for collecting data related to public health. Their proposed algorithm is cost-effective and can also be utilized to identify patients having mild symptoms and indications of COVID-19. In order to obtain relevant results rapidly, the proposed technique should be applied in a well-timed manner. The method developed could also be useful to efficiently identify and control COVID-19 in areas which are quarantined due to the increasing reach of SARS-CoV-2.

Mohammad (Behdad) Jamshidi et.al. [6] explored various Artificial Intelligence and Deep Learning-based approaches to combat the pandemic of COVID-19. Some Deep Learning methods to achieve this include LSTM networks, GANs, and Extreme Machine Learning (ELM). The authors describe how data, structured or unstructured is combined using a bioinformatics procedure to build platforms for researchers. Each platform includes different forms of data, which facilitates performance improvement in real-world applications, such as diagnosing and treating patients infected with COVID-19. They investigated recent medical reports related to building Artificial Neural Network-based tools by using appropriate inputs and required outputs for problems associated with COVID-19.

Sivaramakrishnan Rajaraman et.al. [7] demonstrated the use of a group of iteratively pruned deep learning models for detecting pulmonary manifestations of COVID-19 from CXRs. The CXRs are classified into bacterial pneumonia, absence of any infection, or COVID-19 viral infection. Pretrained ImageNet models and a custom CNN are evaluated on CXR images of patients, which are openly accessible, for learning feature representations. These models are then fine-tuned to enhance performance for classification into the appropriate category. Performance is further improved by a combination of different strategies. The weighted average of pruned models having best performance resulted in an accuracy of 99.01%, with 0.9972 area under the curve for detection of COVID-19. This improved performance was the result of a combination of ensemble learning, model transfer, iterative model pruning, for COVID-19 detection from CXR images.

Muhammad Ilyas et.al. [8] discussed various approaches and challenges involved in detecting COVID-19. With the rapid rate of transmission of the virus on contact, it is necessary to build an automated detection system. ResNet, Googlenet, and many other deep learning models are being used for detection of pneumonia in ailing patients, with the challenge of deducing whether it was caused by COVID-19 or other bacteria or fungus-related infections. The Deep Learning based approach proposed for the detecting COVID-19 showed encouraging performance, with VGG19 exhibiting 98% accuracy, InceptionV3 showing 96%, and ResNET and ResNet50 displaying accuracies of 96% and 95% respectively. Although pneumonia has other causes as well, the above models were trained for a binary classification on 50-100 CXR images of patients diagnosed with COVID-19 as well as those with no such condition.

Abid Haleem et.al. [9] discussed Big Data applications during the COVID-19 pandemic. They state that a lot of information has been created due to the spike in the number of cases and patients' related data, which needs to be stored using various storage technologies. Big Data contributes information related to various criteria to control, detect and prevent the escalation of COVID-19. It helps to identify risk levels by capturing the medical history of infected patients and identifying cases. It can aid the identification of individuals who may have interacted in person with people infected with COVID-19, using a person's travel history. By keeping a record of fever and additional symptoms, it can recommend medical attention, if required. Big Data can also help to identify whether a patient is COVID-19 positive at an early stage, and subsequently predict other individuals who could be infected in the future, by also factoring in the people who enter into an affected area or city. It can manage the fast-moving data related to the disease, created due to the trends in cases and movement of people during the imposed lockdowns. A major contribution of Big Data is the provision of information related to the pandemics which have occurred in the past, which could subsequently fast-track medicine and equipment development to help control COVID-19.

Feng Shi et.al. [10] reviewed the timely actions taken by medical professionals towards imaging of medical data related to COVID-19, powered by AI. Segmentation is an essential procedure in medical images, as it distinguishes various crucial regions, such as lungs, lobes, and other regions and lesions affected, in CXR images. Diagnosis of COVID-19 is a crucial application which requires image segmentation. U-Net, which performed lung segmentation on CT images is used to distinguish COVID-19 cases from pneumonia acquired by community transmission. Image segmentation is also used for quantification, which can serve a vital purpose in applications in the medical field. VB-Net, which is used for segmenting lung, lung infection and lobes of the lungs, can also perform quantitative analysis to indicate signs of COVID-19 progression, along with predicting the severity and visualizing lesion distribution by calculating percentage of infection. A U-Net based algorithm was also created to perform segmentation of lung lesions, which can extract essential features, thus helping to predict the required period of staying in a hospital. Hence, CT images are widely used, along with multiple deep learning methods, to measure lung lesions and delineations, thus helping radiologists to detect lung infections and perform quantitative analysis to identify COVID-19.

Quoc-Viet Pham et.al. [11] surveyed the importance of Artificial Intelligence and Big Data in responding to the COVID-19 pandemic. Some research has been conducted to improve

reverse transcription polymerase chain reaction (RT-PCR) detection technique, a method for classification of respiratory viruses. They proposed a framework based on mobile phones to detect COVID-19 and conduct surveillance. Most individuals across the globe have access to internet connections, as most countries now have wireless connectivity. The Deep Learning model can be trained with the use of cloud computing services, which is later passed to mobile phones for serving the intended purpose. Hence, the authors presented a framework relying on relevant and insightful data for identification of sequences of antibodies to prevent the growth of COVID-19, by using a combination of medical knowledge, Artificial Intelligence and Big Data. The dataset for this purpose initially contained 1831 sequences of antibodies and antigens of viruses like Dengue, Ebola, and H1N1, along with the corresponding half maximal inhibitory concentration (IC50) values. They created the final dataset, containing 1933 samples, by making use of the RCSP protein data bank to obtain 102 samples to append to the initial dataset.

Noviyanti. P et.al. [12] studied the tweets on Twitter by the public, along with the common sentiment related to the outbreak of COVID-19. They performed Social Media Analytics (SMA) by making use of several classification and clustering algorithms to retrieve, store, process and visualize data. For example, the community detection algorithm was used for the detection of communities and their presence on social media. They utilize social media as a source of data for research purposes, in order to conduct various analyses. Google Colab was used for conducting analysis on data obtained from Twitter by making use of an API key. Their research can obtain different types of tweets related to COVID-19 and also perform sentiment analysis on them. The COVID-19 hashtag was used as a filter for retrieving relevant tweets for research purposes. The work done can be further developed to build a more robust system by conducting analysis related to other aspects of COVID-19, which can all be integrated by using data mining methods.

Mohammed Emtiaz Ahmed et.al. [13] presented social media analysis in a journal related to the public's views on reopening. Social media websites are one of the best possible sources, where people disseminate information, as well as share their thoughts. The data for sentiment analysis comprised Twitter posts in the period of May 3 to May 5, 2020. In order to obtain a uniform structure and reduce the noise, they removed all stop words by making use of Python's NLTK stopwords list. Tweets were filtered by the state name in the USA, and tweets having the word 'reopen' were chosen. To perform sentiment analysis, Python's TextBlob library was used, which analysed tweets based on polarity and subjectivity. The IBM Watson Tone Analyzer was used for emotion detection in the tweets. The ToneAnalyzes API detects emotions from a single sentence, hence the data from tweets is passed in sentential format to detect analytical, tentative, confident, joyful, fearing, sad, and angry tones of the tweets. The most frequent words found from the tweets were presented using Word Cloud Representation and N-gram Representation. In the Word Cloud, the size of the words depends upon their frequency of appearance in the data corpus. The n-gram representation, where n is either 1, 2 or 3, where all tweet chunks were merged to form a flat list, where each chunk was of size n. The words 'business', 'economy', 'back', and 'need' in the 1-gram representation show the weakened economy and negative impact on business, along with the necessity of reopening. The words 'social distancing', 'stay home', 'wear mask', and 'contact tracing' in the 2-grams indicate awareness about the pandemic. The optimism about overcoming the loss is clearly indicated by 'the economic breakdown', 'stock market

fluctuation', 'business plan', and 'training course' in 3-grams. The results of Emotion Analysis indicated the tones in the following order: 'Analytical' (34.7%), 'Joy' (17.35%), 'Tentative', 'Sadness', and 'Confident'. From Sentiment Analysis, it was inferred that many tweets had a 'neutral' sentiment, with a 'positive' sentiment being the second-most frequent. From the analysis of tweets, they concluded that the public shows a positive attitude towards reopening, keeping the economy in consideration. However, it also expects organized plans to be implemented by authorities to avoid a potential second wave of COVID-19.

Siddique Latif et.al. [14] presented a systematic analysis of efforts undertaken and challenges faced when making use of data science to fight COVID-19. They summarized key research areas, potential modelling and use cases, where data scientists could provide their contribution for bringing this pandemic under control. Epidemic models, used for observing visible behaviour of infections, can be developed and parameterized. Compartmental models are used to divide people into compartments, while simultaneously observing movement of people across different compartments. Simulation models also have significant applications in fighting the spread of the disease, such as estimating quarantine periods and determining strategies for social distancing. Provision of hospital beds, allocation of resources across various sectors, deciding the prudent time to admit and discharge patients are other use-cases of simulation models. Textual Analysis can be carried out to relate terms in the English vocabulary to the virus, along with the symptoms and protein terminologies. Social media research can also be carried out regarding implementation of government policies. Twitter data from multiple countries in their respective languages can be compiled to identify common responses and the public opinion concerning the COVID-19 pandemic.

R. Sujath et.al. [15] developed a forecasting model for predicting the growth of COVID-19 in India, using Machine Learning algorithms. Various numerical models are being used to predict the growth trend of this pandemic. Various factors influence all these models, with all investigations relying on a potential of an increase in the spread. The authors developed a model for predicting the potential spreading of COVID-19, using linear regression, vector autoregression and multilayer perceptrons. A COVID-19 dataset from Kaggle was used to predict the incidence, distribution and control of COVID-19's escalation throughout India. They also analyzed the possible pattern of effects in India caused by COVID-19. The future state of cases was predicted by using the data related to confirmed cases, deaths, and recoveries across the country for a specific time period. For maintaining the effectiveness of the model, recent data should be added regularly.

Vaibhav Bhatnagar et.al. [16] analyzed the impact of COVID-19 using data specific to patients who either reside in India or had visited India in recent times. The dataset used was sustained by authorities from the government, with continuous updates in a short frame of time. Experiments revealed that the age of an individual did not play a significant role in susceptibility to this disease. There was an observable relationship between the gender and type of transmission; local or from another country. However, the gender of a person and the Indian state where they received treatment from were not related to each other.

3. Reviewed Methods

This section explains various approaches related to detecting COVID-19 from CXR images and CT scans. It also reviews methods related to creating synthetic data, and segmentation of infected regions.

3.1. Deep CNN Covid Net

CovidNet was proposed as a collaboration between designing a network prototype and design exploration by machines, to form a neural network with the purpose of detecting COVID-19 positive patients from their CXR images. A dataset called COVIDx was created, to provide data for the COVID-Net model training and testing purposes. It consists of 13,975 CXR data gathered from around 13,870 people. The network was designed for a three-way classification: normal cases, cases of general viral or bacterial infection, and COVID-19 positive cases. The reasoning for these classifications was to assist medical professionals to recommend PCR testing in suspected COVID-19 cases, and allocate a treatment method depending on the infection (COVID-19 or non-COVID-19) contracted by a given patient.

A machine-driven procedure of exploring alternative designs is undertaken to develop a neural network, with ideal designs developed by following the initial prototype, along with design requirements specifically chosen by humans. Generative synthesis is chosen as the ideal strategy, as it ensures a more pliant design, which incorporates requirements specific to various domains for operation. A generator and an inquisitor module work together to design the most favourable micro and macro architectures, involving long-range connections, in order to satisfy the human-specific expectations from the design of the neural network, which stipulate COVID-19 result sensitivity and value for predicting positives correctly to be 80% each. These requirements are necessary in order to avoid an influx in the number of false-positives, which would be counterproductive to the objective of reducing workload on medical staff.

In order to develop COVID-Net, it is essential to make it a transpicuous system to ensure the good health of patients who are suspected of COVID-19. Therefore, COVID-Net has been qualitatively analyzed to examine factors critical to making decisions, which confirms that they are being taken are based on appropriate information, and not error-prone indicators, such as markup defects. To account for explanations, GSInquire [17], an essential component with respect to the generative synthesis technique, which enabled the creation of COVID-Net, was used to analyze the system qualitatively. GSInquire involves a pair called generator-inquisitor, where both components work together for generation of neural networks, as well as intuitively indications about potential improvements to the neural network architecture.

3.2. Synthetic Data Augmentation using Auxiliary GANS

Next in line is the discussion about augmenting the visual data. The dataset, with a total of 1124 CXR images, comprised 403 infected images, and 721 non-infected images. Due to the fact that COVID-19 is a recent emergence, the number of publicly available CXR images

available for COVID-19 positive cases is significantly lesser than normal CXR images. Due to the less volume of the dataset, generation of synthetic image data was undertaken to further increase the volume of the dataset, and subsequently enhance the performance and accuracy of determining COVID-19 positive cases from the CXR images. A VGG16 architecture contains twelve convolutional layers, with 3 x 3 filters, with some convolutional layers preceding max-pooling layers, and three Dense layers at the termination. This VGG16 network has been used for detecting COVID-19 in patients.

An Auxiliary Classifier Generative Adversarial Network (ACGAN) has been put into use here. Generative Adversarial Networks (GANs) create virtual data, which can be used in real-world scenarios, by using two neural networks, which work in opposition to each other. One type of GAN, Conditional GAN (CGAN), enables the neural network to improve the quality of data, by relying on outside sources of data and other parameters. ACGAN is used to combine the features of a class label and some noise, which are both used by the generator to generate fake images. The dataset consists of a total of 932 CXR images for training, 331 of infected images and 601 of non-infected images. 192 CXR images for testing and verification of overfitting include 72 infected images and 120 non-infected images. Resizing and normalization is performed for preprocessing the images.

3.3. Deep3D Multiple Instance Learning based Pooling

An attention network based deep three-dimensional multiple instance learning (AD3D-MIL) model has been proposed. The input images are transformed into several 3D instances, from which a bag representation is obtained by pooling using an attention network based multiple instance learning. A neural network transforms the obtained bag representation into a Bernoulli distribution. This three-step process is then integrated into a combined 3-dimensional deep neural network, with MIL applied in order to perform multi-class classification.

A deep instance generator is proposed, which generates instances from one CT scan image. This generator is usually a 3D convolutional neural network, which outputs three dimensional maps of features, when a CT scan is passed into it. The input shape is (H x W x S), and the output shape is (H x W x S x D), where H denotes the height and similarly, W, S and D stand for the width, spatial and dimensions of the various features of the maps, in order.

Thereafter, an adaptive MIL pooling would give a better result compared to maximum or mean-based pooling, which would not perform well on data from different sources. Hence, the attention-based MIL is combined with the given AD3D-MIL framework, for diagnosing COVID-19 cases. Attention-based MIL pooling is denoted by the following function.

$$z = \sum_{n=1}^{N} \quad a_n h_n \tag{1}$$

$$a_n = \frac{exp\{w^T tan\, h(V h_n^T)\}}{\sum_{j=1}^{N} \quad exp\{w^T tan\, h(V h_j^T)\}} \tag{2}$$

,

where $w$and V and are trainable neural network parameters, and $\{h_1, h_2, \ldots, h_N\}$are N instances.

Furthermore, the bag labels are transformed into a Bernoulli distribution in the end by two fully-connected functions, which serve as function(s) for transformation of features, hence obtaining a bag label Y_i, which is determined with the help of a threshold value, generally set at 0.5. All bag representation values greater than this threshold value are classified as having a bag label $Y_i = 1$, while the others are classified as $Y_i = 0$.

Finally, an end-to-end optimized model is thus created by combining the generator, attention-based MIL pooling and transformation function.

3.4. COVID_MTNet based diagnosis with multi-task deep learning methods

The tasks of detecting COVID-19 and detecting the region of interest (ROI) for subsequent detection of the infection, with different sources of data such as CXR images and CT scans, have been implemented using different models. Inception Recurrent Residual Neural Network based on the popular CNN InceptionNet has been widely used for tasks involving detection and diagnosis, while another NABLA-N neural network has been applied to segment the region infected by COVID-19 in the images.

For detection of COVID-19 from CXR images, initially, the distinction of pneumonia and normal CXR images is performed by using the IRRCNN model. Then normal CXRs and COVID-19 infected CXRs are used to re-train the model by applying transfer learning, hence diagnosing COVID-19 cases. Thereafter for segmentation of COVID-19 infected regions from CXR images the NABLA-N model was proposed to decipher the CXR images to detect the segment affected by COVID-19. Along with a mathematical approach, an image mask is created, which will exclusively extract the infected region from the images passed in as input.

In order to detect COVID-19 from CT scan images the approach for detecting COVID-19 from CT scans remained similar to the one used for detection from CXR images. The dataset was created by compiling normal as well as COVID-19 CT scan images. Consequently, for segmentation of COVID-19 infected regions from CT scans, a similar approach to the one used for segmented COVID-19 infected regions from CXR images was used. The model was trained on open source data containing two-dimensional lung CT images.

4. Experimental Setup

The architectures and hyperparameters for the neural networks explained in section 3 are explained below.

4.1. DeepCNN CovidNet

Design: COVID-Net uses the network design methodology of residual Projection-Expansion-Projection design, which has the following structure:

i) Input features are projected to a lower dimension, using 11 convolutional layers.
ii) These features are subsequently expanded to higher dimensions by passage through another 11 convolutional layers.
iii) The spatial features are obtained by representation in a depth-wise format, preserving computing power.
iv) Another 11 convolutional layers are added, which obtain features again reduced to a lower dimensionality.
v) The last 11 convolutional layers which obtain the final features, by expansion to a higher dimensionality.

Further, in order to simplify the process of training and enable better representation, while simultaneously limiting the required computing power, long-range connectivity is selectively employed, instead of excessive connections. Densely connected deep neural networks exhibit better results and are easier to train by using long-range connectivity. In COVID-Net the four densely connected layers act as points of connection to enable long-range connections between layers.

The weights for the CovidNet have been obtained by training already on the dataset provided by the ImageNet architecture, followed by training on the COVIDx dataset. The optimizer used was Adam, with the learning rate set to decrease after a certain number of epochs, when the metrics' results reach a plateau. The epochs and batch size were set at 22 and 64 respectively, with the initial learning rate set at 2e-4. The data was augmented by making various enhancements, including intensity adjustments and image flipping.

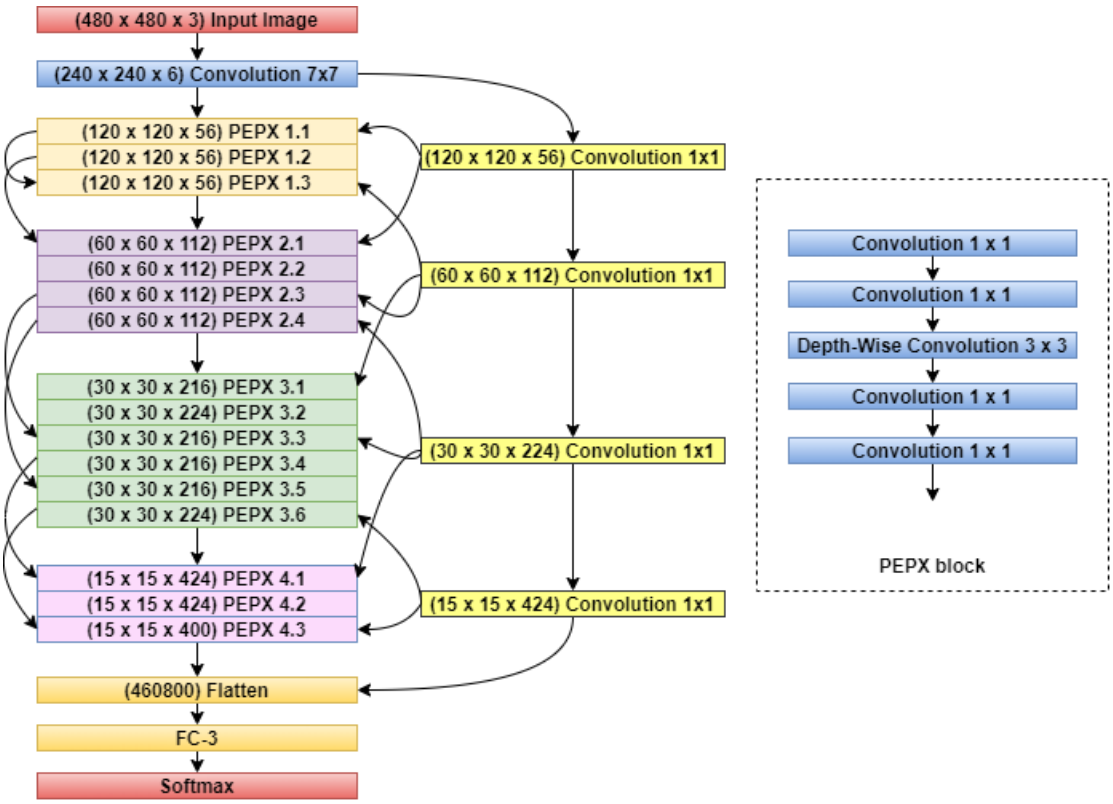

Figure 1. Architectural Description of Covid-Net

4.2. Synthetic Data Augmentation using Auxiliary GANS

As discussed in the Proposed Methodology, CovidGAN consists of two components; a generator and an architecture.

1. Generator:
It passes a class label and noise vector as input, to obtain a single image of dimension 112 X 112 X 3 as output. For obtaining a categorical format input, the label is passed through a layer of embedding of depth 50. This passed through a series of dense layers, with a linear activation, to obtain several low-resolution images. The concatenated class label and noise tensors are passed through four convolutional layers which are transposed, to obtain feature maps of dimensions 112 x 112 x 3. Batch Normalization and ReLU activation are applied to all convolutional layers except the last, which has the hyperbolic tangent function as its activation.

2. Discriminator:
It is an architecture based on Convolutional Neural Networks, in which a 112 X 112 X 3 size image is passed in as the input, to obtain two outputs, one classifying the image as real or fake, and the other classifying the CXR image as COVID-19 positive or negative. Each block contains a layer for batch normalizing preceded by one single convolutional layer and followed by a Leaky ReLU activation function, and then a Dropout layer with rate set to 0.5

to minimize overfitting. The final shape of the input image is 7 x 7 x 512, with the whole network having 2 million trainable parameters in total.

3. Training:

The CovidGAN model is composed of a discriminator followed by a generator. The discriminator weights are set to non-trainable, while the generator can update weights based on the input received from the discriminator. Both non-infected and infected synthetic CXR data is generated by resizing the input to 112 X 112 X 3, and scaling its pixel values to the range [-1,1]. Owing to the minimal memory requirements and efficient computation, Adam is chosen as the optimizer, with momentum set at 0.5. The first discriminator output layer uses binary_crossentropy as the loss function, and the second output layer uses sparse_categorical_crossentropy for computing the loss. A batch size of 64 is chosen, along with learning rate = 0.0002, to initialize the training, with twenty-four million parameters.

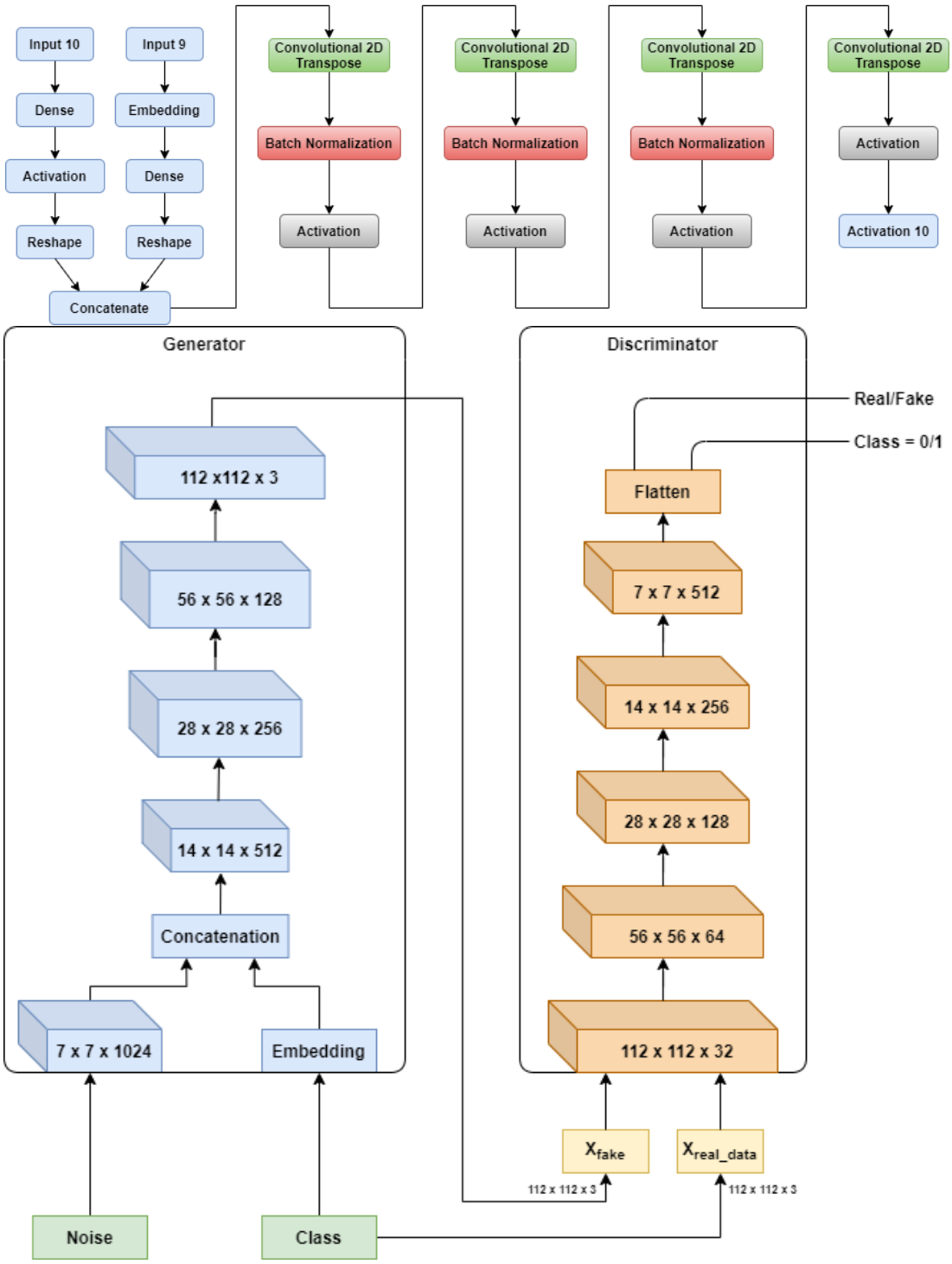

Figure 2. Description of CNN and ACGAN used in CovidGAN

4.3. Deep3D Multiple Instance Learning based Pooling

The dataset contained 460 CT scan images, including 230 CT images of 79 COVID-19 positive individuals, 130 images of people with no disease or infection, and 100 images of patients infected with pneumonia. For splitting the data into training and test sets, no CT scans of the same individual can be present in both sets at the same time.
Common pneumonia patients include bacteria-affected as well as viral-infected individuals. The individuals having absence of pneumonia could be affected with diseases other than COVID-19 and pneumonia. Only those images which included an icon were added. Common pneumonia images may increase the complexity to correctly decipher COVID-19 positive cases, but they were essential to create a distinction between COVID-19 and pneumonia.

The AD3D-MIL algorithm is executed twice, initially by classifying a case as COVID-19 positive or not. The latter training session effectively assessed three classes, usual CT scans, COVID-19 infected CT scans, CT Scans with common-pneumonia. The division for training, testing and validation was of the ratio 60%, 20% and 20%. This was further combined alongside five-fold cross validation. To verify genuineness of results, each of the above sets were implemented five times. The main metrics used for assessment included calculation of F1 score, accuracy, precision, recall, and comparison of performances of C3D and DeCoVNet models with the AD3D-MIL algorithm. The input shape was set at 256 x 256, with variable number of slices chosen. Traditional methods of augmenting data to increase performance metrics were performed on the presented dataset. The learning rate has been set to 1e-5 before training, along with the Adam optimiser chosen for the network. The models were then trained for a total of 100 epochs.

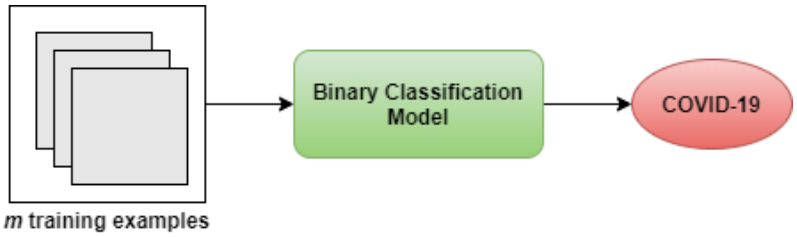

Figure 3. Regular Supervised Learning

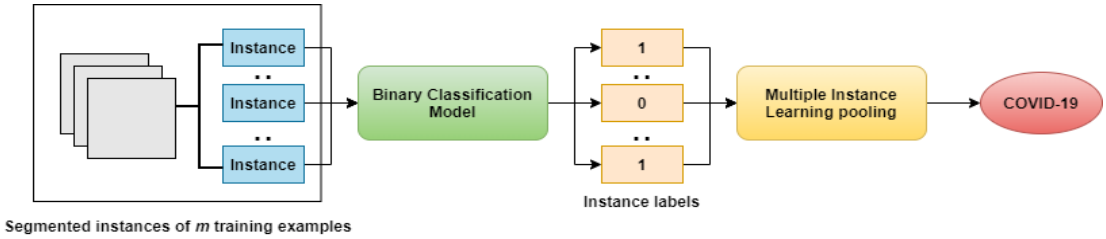

Figure 4. Regular Multi-Instance Based Learning

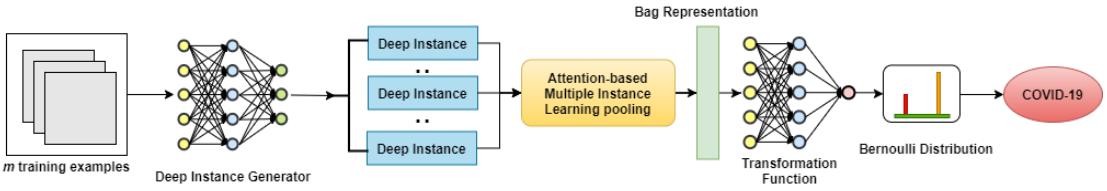

Figure 5. Stated Attention Based Multiple Instance Learning

4.4. COVID_MTNet based diagnosis with multi-task deep learning methods

To implement the various neural networks for detecting COVID-19 as well as segmenting the infected areas, a dataset compiled from various sources containing 5216 samples was used. Out of those, 1341 images denoted normal CXRs, and 3875 images denoted pneumonia infected CXRs, with all the input images resized to 128 x 128, owing to computational limitations. Due to the vast difference in the number of samples in each category, various data augmentation methods were used.

To perform segmentation of images for locating infected regions, 704 CXR images were gathered along with masks to center the infected area. The pixel sizes of the images were resized to obtain dimensionality of 192 x 192 x 3. A drawback of this procedure is slight information loss in the data. 80% of the data is used for training the model, while the remaining 20% is used for validation purposes.

5. Results and Discussion

5.1. DeepCNN CovidNet

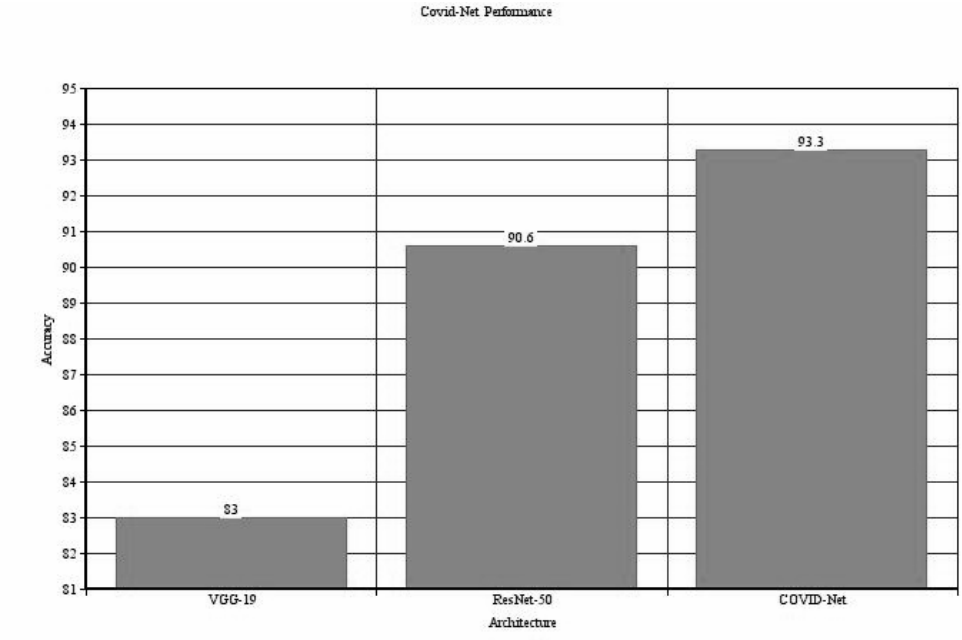

Figure 6: Accuracy of CovidNet

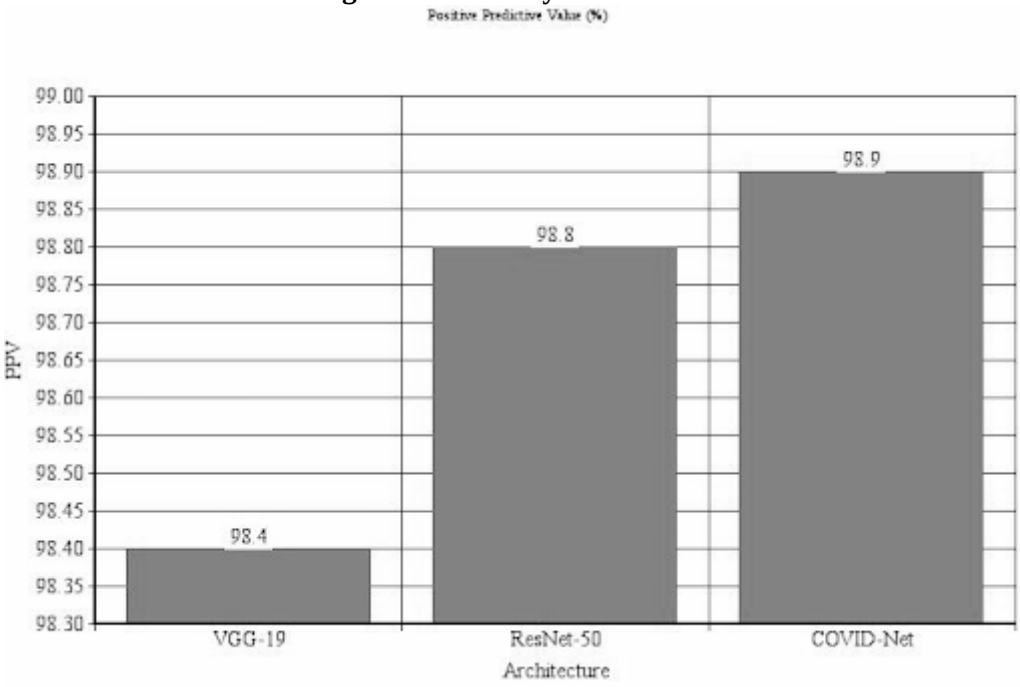

Figure 7: Positive Prediction Value of CovidNet

The accuracy on the testing set, sensitivity, and positive predictive value (PPV), are the metrics used to evaluate the COVIDx dataset. COVID-Net performs well, with an accuracy of 93.3%. This performance validates the reasoning for a design strategy involving human-

machine collaboration, which creates new neural network architectures with customizations, by taking into account the task it is meant for, and the data available for training purposes. This strategy is ideal for detecting diseases, as the available data is continuously evolving with time, and subsequent changes in neural network architectures are essential. With a PPV value of 98.9%, Covid-Net mostly avoided the issue of incorrect positive detection. A high PPV value is completely necessary, since significant false positive detections would essentially harm the purpose of creating a neural network to assist medical professionals in testing. In order to effectively generalise across various circumstances, the methodology for training can be improved, accompanied by an increase in the amount of data used to train the model.

5.2. Synthetic Data Augmentation using Auxiliary GANS

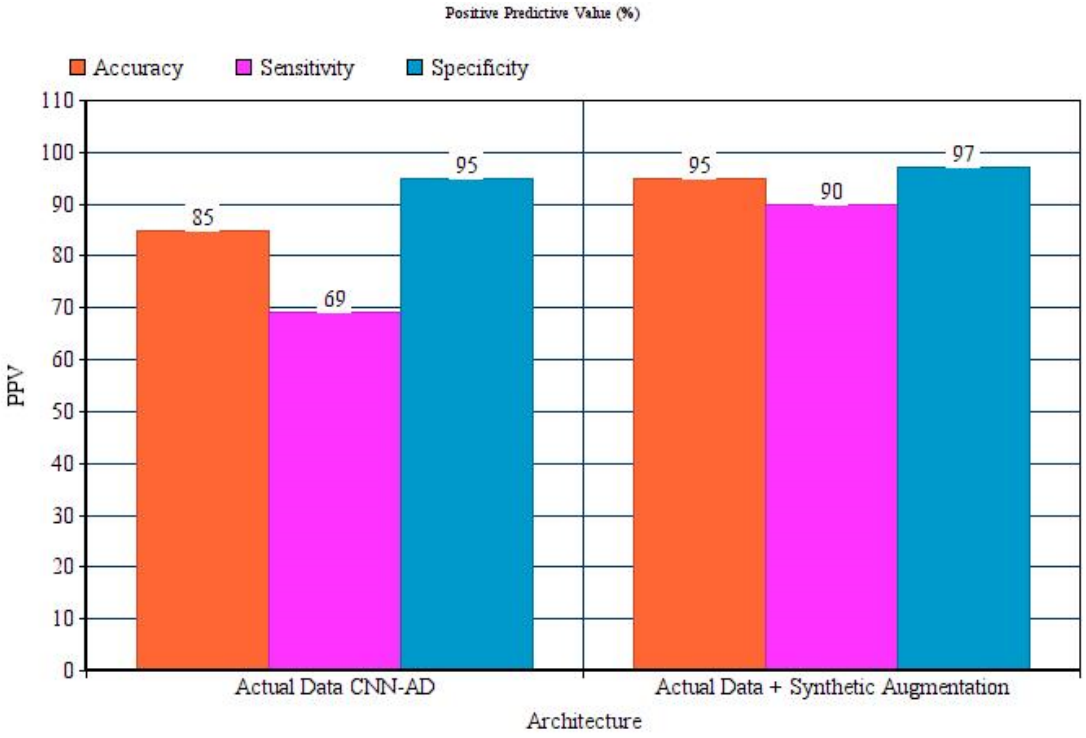

Figure 8: Comparison of Actual Data Metrics with Combined Actual and Synthesised Data Metrics

For judging the efficacies of CovidGAN, the performance of neural networks using the synthetic data created by it was analyzed. A VGG-16 CNN architecture was used to detect COVID-19 in individuals. The testing samples contained 72 COVID-CXR images, and 120 non-COVID-CXR images, amounting to 192 samples. When VGG-16 was trained with actual data, an accuracy of 85% was achieved. This increased to 95% when trained using synthetic data produced by CovidGAN. The precision, recall and specificity for COVID positive data were 0.89, 0.69 and 0.95 respectively, which increased to 0.96 precision, 0.90 recall and 0.97 specificity when synthetic data was used. The non-COVID data also exhibited precision and recall of 0.94 and 0.97 respectively. Hence, it can be conclusively determined that synthetic

data creation by augmentation techniques was able to enhance the VGG-16 performance for diagnosing the presence of infection from visual CXR data.

5.3. Deep3D Multiple Instance Learning based Pooling

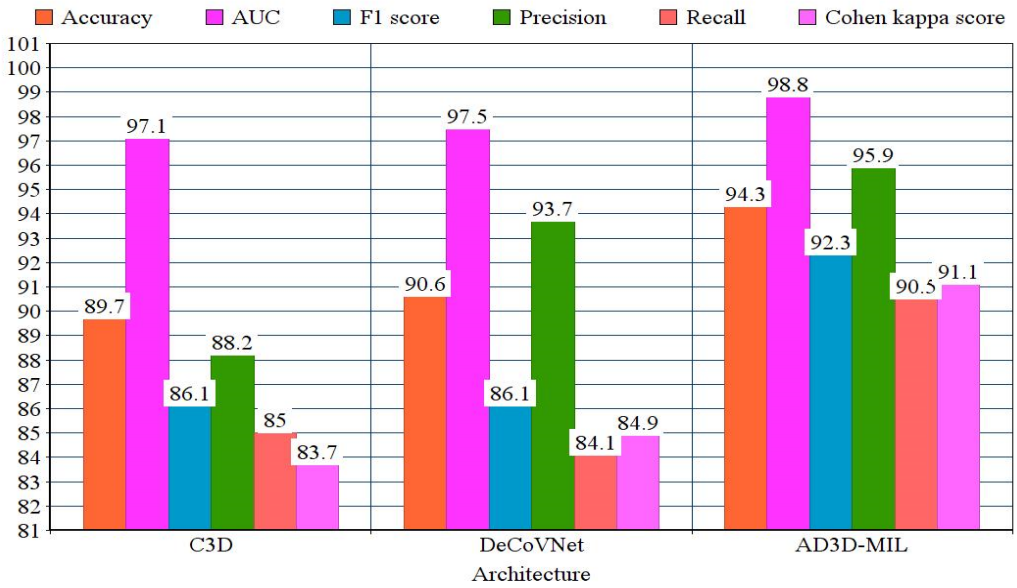

Figure 9: Comparison of various metrics for C3D, DeCovNet and AD3D-MILNetworks

The speciality of the AD3D-MIL algorithm lies in its superiority over other algorithms like C3D and DeCovNet by obtaining a classification accuracy of over 94.3%. It successfully overcomes the dimensional nuances of CT scans and irregular annotations and the confusion matrix signifies that AD3D-MIL obtains minimal false positive errors and most probably diagnoses COVID-19 with minimum incorrect negatives. Verification of the experimental results is done using statistical analysis to test statistical relevance. T-tests help juxtapose the AD3D-MIL alongside DeCovNet, displaying over a marginal five percent level of significance and a corresponding 0.008 p-value. The approbity of the aforementioned AD3D-MIL pooling is shown, compared to DeCovNet. It is also shown that the p-values of all juxtaposed methods are well below 0.05. Based on these analyses it can be successfully concluded that the idea of superimposing a CXR screening as an MIL problem is justified.

5.4. COVID_MTNet based diagnosis with multi-task deep learning methods

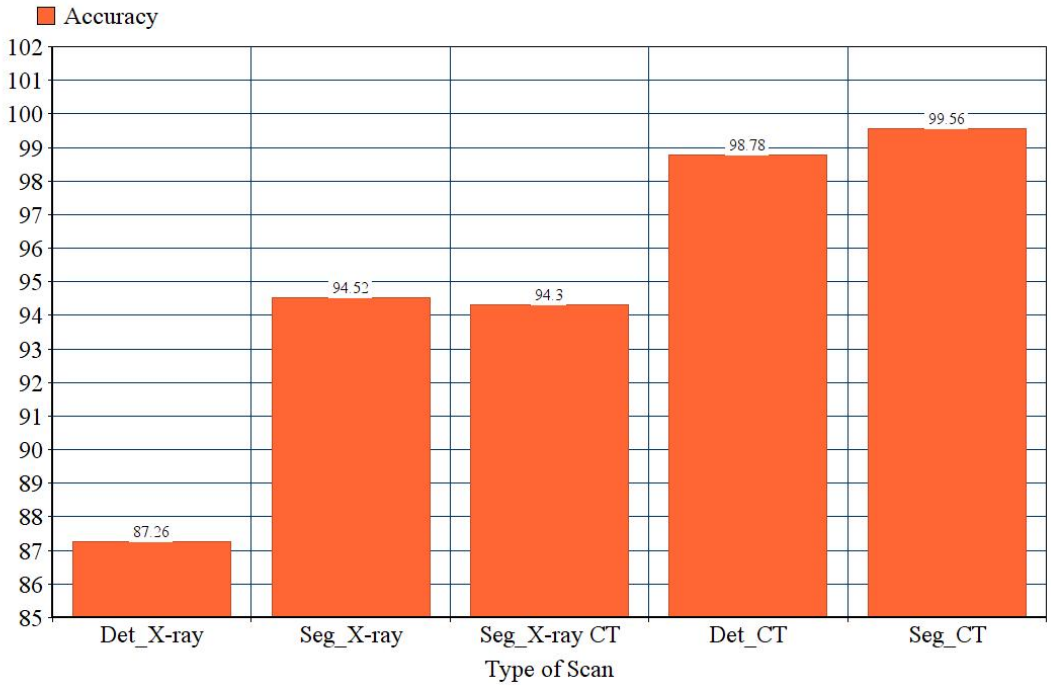

Figure 10: Comparison of Training accuracies of Covid_MTNet for various types of scans

CXR images have been the most widespread source for detection of COVID-19 with a testing accuracy of 87.26% observed for detecting pneumonia, from 624 CXR images, including 390 pneumonia samples. When the IRRCNN model was used for detecting COVID-19 (with initial weights equal to the pretrained weights from training on pneumonia detection), an accuracy of 84.67% was achieved, for 67 new samples. In order to segment the infected region, 57 new CXR images were used. The global accuracy and F1-score were 0.945 and 0.946 respectively.

CT scans are a more reliable source for determining COVID-19 cases, thus assisting doctors effectively. The testing accuracy for 45 CT scan images was observed as 98.78%. The trained NABLA-N model was used for detecting the infected segment. The global accuracy and F1-score were 0.9956 and 0.9885 respectively.

The aforementioned method for pneumonia diagnosis gave nearly 87.26% testing accuracy, proclaiming the IRRCNN as better than the other architectures. The proposed methods use

patch based detection leaving it vulnerable to infected region extractions where there are high chances of falsely detecting a positive patient negative and vice versa. Comparatively, the paper suggested a pixel based analysis for the infected region using segmentation. The quantitative and effective results shown by these methods are demonstrated by the results of the affected regions significantly reducing false results.

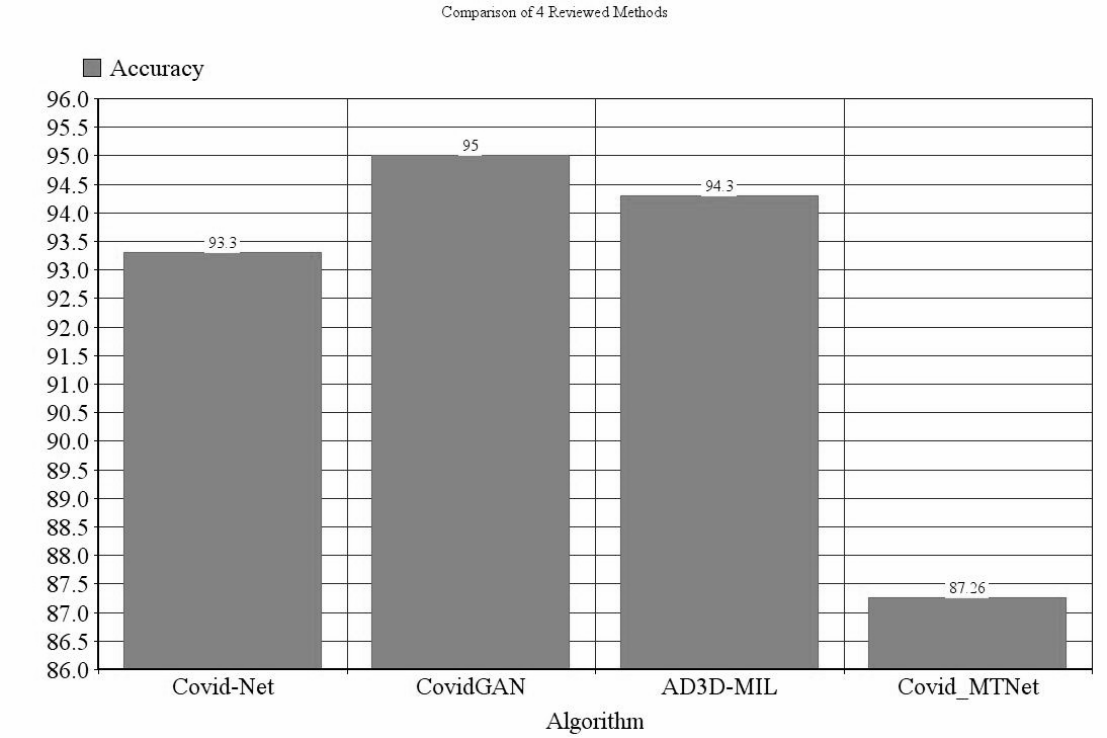

Figure 11: Comparison of the testing accuracies of the four reviewed algorithms for different sample sizes

6. Conclusion

The four methods that have been discussed above, have been created on an acute time deadline and on very scant datasets, hence all of them have achieved remarkable results and verification from minimal resources. Based on the amount of data available and the precision and accuracy generated by the corresponding models, it can be concluded that CovidGAN using Auxiliary
Classifier GANs is the preferred model for screening, segmentation and corresponding detection for Chest X-Ray in order to positively judge the given scan images as Covid-19 affected.
In this paper, the authors have presented an all-inclusive examination and evaluation
of the CXR images for Covid characterisation using a VGG16 architecture. Its efficacy is enhanced when the want for a good dataset is satisfied using synthetic collection of lung chest X-rays using GAN networks. This helps prevent the problem of overfitting and reduced effectiveness of the architecture to effectively judge a variety of images successfully. This innovative algorithm has evidently outperformed other algorithms discussed and requires a regular VGG16 CNN thus carrying out effective computations. Although other methods have achieved higher accuracy than CovidGAN for Segmented Chest X-Rays, the number of samples considered is far lesser and we safely assume that for the sheer volume of the dataset

populated by CovidGAN, the accuracy achieved is the most preferable by far. The task of detecting the presence of the virus in a variety of CXR images, further enhanced by the CovidGAN has been achieved with a high specificity and is a reliable tool for potential use in detection of COVID-19 in patients and thus categorizing which patients are most necessarily required to undergo a PCR Test for complete surety.

7. Future Scope

The major challenge surrounding the neural networks built for prediction, forecasting and mainly screening of CXR images for effective diagnosis is not just the lack of data but also the lack of diverse data. The presence of less than two thousand images of CXR and other data poses some serious problems such as the overfitting of the model to the given data, inability to work effectively for a slight variation in the type of images. Hence with further developments, several advancements and scientific and biological unfoldings in the future, it will be potentially easier and more beneficial for several models. The extensive availability of large and intensive datasets along with knowledge of the kind of methods such as augmentation and segmentation that can be applied is sure to open up the avenues for more accurate and robust models. Not only will it make the detection and forecasting of Covid much more advanced and reliable, but also open up the doors for transforming such knowledge for any future disasters of this sort. Since the situation is unprecedented, there is not much knowledge of the methods that will work in the scenario. Hence it is by experimenting different methods on top of these already built architectures to find out the best possible combinations of layers and functions and various additional architectures that ensure the best accuracy. Hence the future scope of this project lies in improving the present methods, ensuring better precision, recall and avoiding false positive diagnoses, making a large dataset available to the public for processing the images better and making these models versatile enough to be used in the future for any pandemics of this sort. Furthermore, as biological research advances and the genomic combinations of the virus and the procedure of breaking its sequence becomes clearer, it opens up the avenues for using Neuro Linguistic Programming for determining RNA sequences that serve as a good cure for the virus.

References

[1] Wang, Linda, and Alexander Wong. "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images." arXiv preprint arXiv:2003.09871 (2020).

[2] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman and P. R. Pinheiro, "CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection," in IEEE Access, vol. 8, pp. 91916-91923, 2020, doi: 10.1109/ACCESS.2020.2994762.

[3] Z. Han et al., "Accurate Screening of COVID-19 Using Attention-Based Deep 3D Multiple Instance Learning," in IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2584-2594, Aug. 2020, doi: 10.1109/TMI.2020.2996256.

[4] Alom, Md. Zahangir & Rahman, M M Shaifur & Nasrin, Mst & Taha, Tarek & Asari, Vijayan, "COVID_MTNet: COVID-19 Detection with Multi-Task Deep Learning Approaches", 2020.

[5] Arni S. R. Srinivasa Rao and Jose A. Vazquez, "Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone–based survey when cities and towns are under quarantine", Infection Control & Hospital Epidemiology, The Society for Healthcare Epidemiology of America,
[6] M. B. Jamshidi et al., "Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment," in IEEE Access, vol. 8, pp. 109581-109595, 2020, doi: 10.1109/ACCESS.2020.3001973.
[7] Sivaramakrishnan Rajaraman, Jenifer Siegelman, Philip O. Alderson, Lucas S. Folio, Les R. Folio, Sameer K. Antani, "Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-Rays", in IEEE Access, 2020, doi: 10.1109/ACCESS.2020.3003810
[8] Muhammad Ilyas, Hina Rehman, Amine Nait-ali, "Detection of Covid-19 From Chest X-ray Images Using Artificial Intelligence: An Early Review", arXiv:2004.05436 [eess.IV].
[9] Haleem, Abid & Javaid, Mohd & Khan, Ibrahim & Vaishya, Raju, "Significant Applications of Big Data in COVID-19 Pandemic", Indian journal of orthopaedics, 2020, doi:1-3. 10.1007/s43465-020-00129-z.
[10] F. Shi et al., "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19," in IEEE Reviews in Biomedical Engineering, 2020, doi: 10.1109/RBME.2020.2987975.
[11] Q. Pham, D. C. Nguyen, T. Huynh-The, W. Hwang and P. N. Pathirana, "Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts," in IEEE Access, vol. 8, pp. 130820-130839, 2020, doi: 10.1109/ACCESS.2020.3009328.
[12] Arianto, Dian & Pui, Noviyanti, "Social Media Analysis: Utilization of Social Media Data for research on COVID-19", 2020.
[13] Ahmed, Mohammed & Rabin, Md & Chowdhury, Farah, "COVID-19: Social Media Sentiment Analysis on Reopening", 2020.
[14] Latif, Siddique & Usman, Muhammad & Manzoor, Sanaullah & Iqbal, Waleed & Qadir, Junaid & Tyson, Gareth & Castro, Ignacio & Razi, Adeel & Boulos, Maged & Weller, Adrian & Crowcroft, Jon, "Leveraging Data Science To Combat COVID-19: A Comprehensive Review", 2020, doi: 10.36227/techrxiv.12212516.
[15] Radha, Suja & Chatterjee, Jyotir & Hassanien, Aboul Ella, "A machine learning forecasting model for COVID-19 pandemic in India", Stochastic Environmental Research and Risk Assessment, 2020, doi: 10.1007/s00477-020-01843-8.
[16] Vaibhav Bhatnagar , Ramesh Chandra Poonia , Pankaj Nagar , Sandeep Kumar , Vijander Singh , Linesh Raja & Pranav Dass (2020): Descriptive analysis of COVID-19 patients in the context of India, Journal of Interdisciplinary Mathematics, DOI: 10.1080/09720502.2020.1761635
[17] Lin, Zhong & Shafiee, Mohammad Javad & Bochkarev, Stanislav & Jules, Michael & Wang, Xiao & Wong, Alexander, "Do Explanations Reflect Decisions? A Machine-centric Strategy to Quantify the Performance of Explainability Algorithms", 2019, arXiv preprint arXiv:1910.07387.