I. INTRODUCTION

According to the National Institute of Mental Health (NIMH), schizophrenia is a mental disorder that disrupts thinking processes, perceptions, emotional reactivity, and social relationships (McCutcheon, 2019). Typically manifesting in late adolescence or early adulthood, schizophrenia is frequently examined from a developmental standpoint. The latter stage of the condition is characterized by chronic manifestations of numerous symptoms, including cognitive impairment and odd behavior in children (Reis, 2019). This pattern may be a result of both disturbances in brain development and environmental variables, such as prenatal or early life stress. According to statistics collected by the Institute of Health Metrics and Evaluation (IHME), schizophrenia affects roughly 24 million individuals, or 0.32 percent of the global population. This rate affects 1 in 222 adults or 0.45 percent.

The diagnosis for many psychiatric disorders such as schizophrenia is mostly based on patient interviews, which symptoms they exhibited, and whether there is presence or absence of representative behavioral indication (Oh et al., 2019). Ter-Stepanyan et al. (2021) said that when a psychiatrist talks to a patient to figure out what's wrong, they will also do a clinical diagnostic, which could take up to two years if the protocols are followed. The paper also mention because of the extended time frame, as well as the necessity for precision and accuracy in the diagnosis, alternative procedures, such as neuroimaging techniques, are more appropriate.

Another method that could be proposed to replace the conventional way is the brain computer interface (BCI) system. BCI is a system that monitors central nervous system activity or brainwave signals, turns them into artificial output that enhances and improves natural central nervous system output, and so modifies the continuous interactions between the central nervous system and its environment (Wolpaw, 2012).

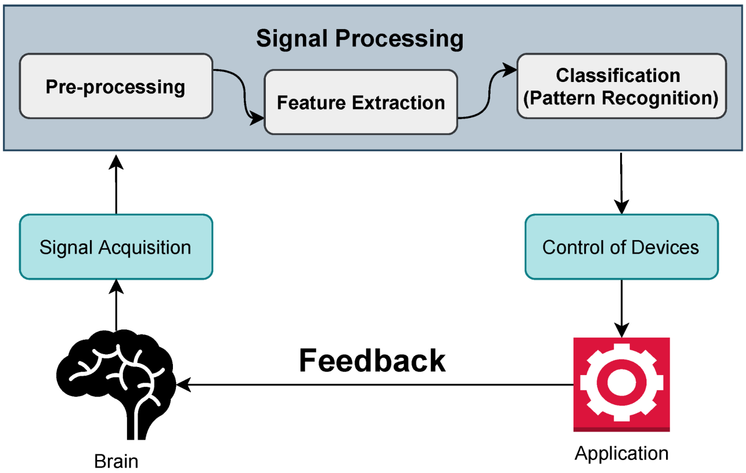

Figure 1. The basic architecture of BCI system (Mridha et al., 2021).

Figure 1, shows how the BCI system is put together from the way of signal extraction from the brain until the process of making it into a workable command. A study by Mridha et al., (2021), the first step was called signal acquisition, which was the process of converting signals that assess brain activity into commands that would be preprocessed further. The second step would be pre-processing the signals that have been gained from the patient. In most instances, the obtained brain signals are noisy and degraded by artifacts. This step aids in artifact removal with the usage of various filtering techniques. The next step is called feature extraction, which combines signal analysis and data extraction. Due to the complexity of the brain activity signals, it was difficult to extract usable information from them. Consequently, processing techniques that enable the extraction of brain characteristics, such as a person's purpose, are required. The signals in the classification step are then classified using artifact-free classification algorithms. The categorization assists in determining the sort of mental job or command the individuals are undertaking. From then on, a command is sent to the feedback device or application during the categorization process. It may be a computer, where the signal is used to move a cursor, or it could be a robotic arm, where the signal is used to move the arm.

Recent advances in BCI research have led to the availability of real-time, advanced signal processing technologies, a greater knowledge of the features and applications of brain signals, and even rehabilitation techniques for motor dysfunction impairment (Daly, 2015). The primary purpose of the BCI system was to aid motor dysfunction within the body by employing external devices to record the user's brain activity in order to regulate it (Birbaumer, 2007). EEG is one of the several ways in which brain signals may be obtained. With the manufacturer-supplied software development kit, these EEG headsets were able to capture expressive data. It has also been used to enable severely crippled individuals to operate an external device using small movements, such as neck motion and blinking (Ruiz, 2013). Numerous doctors and clinician-scientists must be made aware of the advancement of BCI. Due to their applicability in clinical delivery environments and their precision in identifying patients' complaints, these systems may aid in enhancing rehabilitation techniques (Daly, 2015).

The financial weight of medication comes with the passage of time. Although neuroimaging techniques such as ECG (Electrocardiogram) and electromyogram have been used to diagnose schizophrenia, it has been discovered that these models do not yield satisfactory findings. A lack of defined methodologies and approaches has impeded the performance of early identification of schizophrenia (Khare et al., 2021).

Despite the progress in understanding the neural substrates of schizophrenia, current treatments mainly based on antipsychotics do not incorporate existing know-how in etiology and pathogenesis of the disease, such as, the abnormal connectivity phenomenon (Ruiz, 2013). In fact, to date there are no treatment methods that can produce a direct and significant change in brain connectivity, or are specifically designed to tackle this problem. If the early symptoms are misdiagnosed or if the initially prescribed drug or therapy fails to take effect, the best time to control and treat the disease can be lost and it may even cause unwanted side effects to the patient (Charlotte, 2014).

An alternative to this problem is EEG or electroencephalogram, which is one of the most practical and inexpensive functional neuroimaging modalities. Within this modality, it characterizes the electrical activities of the brain recorded from the scalp of the surface of the head with a high temporal resolution and an appropriate spatial resolution (Murashko, 2019),. This modality is crucial for the early diagnosis and detection of schizophrenia by any computerized means. Computer-assisted diagnostic tools for clinical use and research into disease pathophysiology can be developed using machine learning (ML), according to Sun et al. (2021). To deal with the enormous complexity of EEG information, machine learning has revolutionized schizophrenia research. For EEG analysis, the classic machine learning technology (ML) (i.e., a non-deep learning (DL) algorithm) has been the method of choice in recent years, and it has been integrated with a variety of feature extraction methods. DL algorithms have been widely used in medical image and signal processing, and they have been demonstrated to have a lot of research promise. In the vast majority of circumstances, their performance outperforms that of classic machine learning techniques.

Convolutional Neural Network, or CNN for short, is one of the deep learning algorithms that would be employed in this study to evaluate whether the patient has schizophrenia. When it comes to classifying two-dimensional (2D) images, according to Phang et al. (2020), it performs admirably because of its superior capacity to represent spatial patterns. With their sparse localized kernels, CNNs may efficiently train a hierarchy of latent representations that are invariant to tiny translations of the inputs and allow for parameter sharing. There were many benefits in using CNN as the main deep learning which were the weight sharing feature which reduces the number of trainable network parameters which helps in enhancing the large scale network generalization and implementation in avoiding overfitting (Alzubaidi, 2021),. Another reason was because of their feature extraction and classification layers which cause the model output to be organized and reliant to other extracted features.

The learning process of CNN involves in the selection of learning algorithm, in this case its the optimizer, and the employment of several upgrades in conjunction with the optimizer in improving the model output. The primary aim of optimizers were to minimize the error rate, based on multiple parameters such as the loss and accuracy. In minimizing the error rate, the parameters should be updated throughout all training epochs and network should also look for locally optimum answer across all training epochs (Alzubaidi, (2021). The training epoch comprises a complete repetition of the parameter update involving the entire dataset at once. Although it is hyper-parameter, the selection of learning rate need to be done carefully to prevent the imperfection of the learning process.

The main objective of this study is to figure out which optimizer in the CNN model does the best job of identifying schizophrenia patients with a high degree of accuracy and precision. Due to the different results between each different optimizer, a suitable optimizer is crucial in determining which optimizer is best paired with the CNN model. Within the study, a comparison between optimizers was needed to be done to determine the EEG-based schizophrenia dataset in classification using CNN.

II. LITERATURE REVIEW

A. Schizophrenia clinical diagnosis

A psychiatrist performs a clinical interview, examines the patient's past history, and administers psychological tests to assess whether or not a patient has schizophrenia (Prabhakar et al., (2020). Prabhakar mentions that after roughly a month of observation, the assessment procedure began with recognizing the patient's symptoms. Prabhakar found that the continuing evidence of symptoms are then followed for at least six months. Within the paper, positive and negative symptoms are distinguished. Those positive symptoms include delusions, hallucinations, disordered cognition, and disorganized conduct. Anhedonia (loss of interest in many things), avolition (lack of motivation), flat affect (lack of emotion), and alogia (lack of thought) are all negative symptoms mentioned within the paper. Prabhakar founds that schizophrenia does not occur while a person is under the influence of drugs or medicines. In the paper, it mentions that when a patient has a history of autism or another pervasive developmental disability, schizophrenia is only diagnosed if persistent delusions or hallucinations have also been present for at least a month (or less if the patient has been successfully treated).

A complete clinical evaluations were performed at baseline, 6 weeks, 12 months, and 36 months, and protocol clinical interviews (Pelayo-Terán et al., 2018). Pelayo mentions that the evaluations depends on The Scale for the Assessment of Positive Symptoms (SAPS) scale was used at regular intervals for 36 months after the initial presentation to determine the duration of psychosis after treatment (DAT). Within the paper it mentions that the assessment for the patients each week, subscale scores for hallucinations, delusions, bizarre behavior, and positive formal thought disorder were calculated prospectively to determine the severity, duration, and course of clinical symptomatology. Those patients were monitored in the clinic and had quick and easy access to a clinical appointment at any time if any signs or symptoms of clinical exacerbation appeared.

B. EEG signals in schizophrenia patients.

Multimodal imaging in neuroimaging, may be utilized in many additional ways to detect early indications of schizophrenia (Alzubaidi, 2021). Some modalities mentioned within the paper are such as positron emission tomography, functional magnetic resonance imaging, and diffusion tensor imaging are a few of the modalities available. Using a mix of the approaches listed above can be helpful when just one imaging modality is able to reveal the patient's neurological illness (Jack et al., 2009). Combining imaging equipment may not only be too expensive to implement, but the fusion of pictures from two distinct devices might result in motion artifacts (Wehrl et al., 2017). It is therefore necessary to develop a more cost-effective way for diagnosing schizophrenia. Electrical activity in the brain may be measured using electroencephalograms (EEGs), which are signals captured from the scalp. Oh et al., (2019) also mentions that automated detection of neurological illnesses such as epilepsy and depression, Parkinson's, Alzheimer's disease, and dementia would be extremely beneficial.

EEG will be used as an aid alternative for the long-awaited clinical assessments. EEG worked by inserting electrodes into the surface of the scalp to record and describe the electrical activity in the human brain, which is then analyzed (Ahmadlou, 2012). Based on Sun et al., (2021), in addition to epilepsy and Alzheimer's disease (AD), electroencephalography has also been used to study and diagnose schizophrenia and other mental illnesses. Sun found that high-dimensionality and big data volumes make it difficult to directly interpret EEG signals' complex content. Therefore, more approaches for retrieving relevant data from the brain require to be explored and trialed for further analysis. Feature extraction can be used to examine EEG data. Sun mentions that in order to reduce the size of the individual characteristics without losing any of their physical relevance, only the most significant ones from a large number of original signals were selected for further analysis and processing. Sun also finds that for the study of brain state changes in schizophrenia patients, EEG data may be quantified using several feature extraction methodologies, such as those for time-domain features and frequency-domain features by utilizing the algorithms and functions of deep learning.

Deep learning algorithms can automatically take important features from the data and classify them. These deep learning methods imitate the workings of the human brain in data processing and generating patterns of decision-making usage. Many recent developments in neural network architecture design and training have enabled researchers to solve previously intractable learning tasks of deep learning methods (Shalbaf, 2020). As a result, several research works have focused on the application of deep learning as the state-of-the-art in machine learning especially the Convolutional neural network (CNN) in a wide range of computer vision studies especially in medical applications and also for processing EEG signals with very success. Below paragraph are some of the recent research works that are employed in detecting schizophrenia within the patients by utilizing the usage of CNN and EEG signals.

C. Using convolutional neural networks (CNN) to classify the patients using different optimizers

An automated technique was used in a study for diagnosing schizophrenia patients utilizing CNN from EEG data. They gained the data from 14 normal patients and 14 schizophrenia individuals and then employed CNN to distinguish between those patients (Oh, 2019).

EEG recordings were used in a study to examine 2D temporal and frequency domain connectivity aspects, as well as 1D intricate network properties (Phang, 2020). The Multi-domain connectome CNN model within the paper was then used to generate feature maps, which benefited the classification process. For the identification of schizophrenia patients, the model employs several fusion algorithms. In order to determine which algorithm works best, the model with the highest accuracy was chosen as the appropriate algorithm usage.

Another convolutional neural network (CNN) model have been researched for use in schizophrenia diagnosis using EEG data by certain researchers. A study by Shoeibi et al., (2021), said that CNN-recurrent neural network (RNN) models are a subset of Deep Learning networks that are well-known for their capacity to identify various brain illnesses using EEG signals. Their work is proposed by the usage of 1D-CNN-LSTMs (1-dimensional Convolutional Neural Network with long short-term memory) in which they were able to produce encouraging findings.

An optimizer must be chosen first before executing a model of CNN (Zhang, 2018). Zhang mentions that the purpose was to minimize an objective function (often referred to as the loss), which is the difference between the anticipated and expected values. The minimizing process entails identifying the set of design parameters that produce the greatest outcomes in tasks like classification, prediction, and clustering.

Learning rate optimization is one of the approaches used to improve the CNN model's performance (Sevli, 2020). Sevli mentions that the learning rate is a coefficient that is used to update network parameters based on the amount of error that happens during the network's learning process. If the learning rate was too slow, network parameters were changed in extremely small stages, taking a long amount of time to complete. Due to the rapid learning rate, the network might miss the optimum point for minimizing error. As a result, learning rate optimization was very crucial.

Throughout the research, six different optimizers will be employed. These are the optimizers:

1. Adam

Adaptive moment estimation is one of the most often used optimization techniques that calculate the adaptive learning rate for each parameter. It is a computationally efficient stochastic optimization algorithm that incorporates first-order gradients and requires little memory (Zhang, 2018). Adam is often used in the context of machine learning problems involving high-dimensional parameter spaces and enormous data sets that compute learning rates for different parameters using approximations that include first- and second-order moments. Based on Yaqub, (2020), Adam decreases computational cost, implements with less memory, and is invariant to diagonal gradient rescaling. This addresses difficulties such as massive datasets, hyperparameters, noisy data, insufficient gradients, and nonstationary problems requiring little adjustment.

2. SGD

Stochastic Gradient Descent (SGD) is one of the most used deep learning algorithms. It performs parameter adjustments on each training sample (Poojary, 2019). Additionally, SGD performs redundant computations for big datasets, resulting in frequent updates with substantial volatility (Alzubaidi, 2021). This may significantly alter the objective function. Algorithms based on SGD allow users to customize measures inside the algorithm while managing massive datasets (Yaqub, 2020).

3. RMSprop

Root Mean Square Propagation provides parameter learning rates that are tailored to the weight's average recent gradient magnitudes (Postalcıoğlu, 2020). This algorithm is closely connected to the momentum-based gradient descent algorithm, according to Yaqub (2019). RMSProp attempts to overcome the rapidly dropping learning rates of Adagrad by utilizing a moving average of the squared gradient, which normalizes the gradient based on the size of recent gradient descents. Therefore, when the learning rate increases, the algorithm will progress horizontally with larger steps converging more rapidly. The vertical oscillations are controlled by the RMSprop optimizer. Consequently, we may enhance our learning rate, and our algorithm might take greater horizontal steps while converging more rapidly.

4. Adagrad

Adaptive Gradient Algorithm is a gradient-based optimization algorithm. It is an optimizer that modifies the learning rate based on certain dataset properties (Taqi, 2018). Adagrad modifies the learning rate based on the parameters, with larger updates for inconsistent parameters and fewer updates for consecutive parameters (Kingma, 2014). It also eliminates the need to manually adjust the learning rate; nevertheless, its collection of squared gradients in the denominator leads the learning rate to decrease and moderates the interleaving speed (Yaqub, 2020).

5. Adamax

It is an optimizer based on adaptive approximation of low-order moments and is frequently referred to as a modification of the Adam optimizer (Chauhan, 2021). As a result, the Adamax optimizer tends to have less faults than Adam optimizer (Aghdam, 2019). Additionally, the Adamax optimizer is the most stable optimizer with great outcomes and minimum variation across different learning rates (Kandell, 2020).

6. Adadelta

Adaptive Delta is an optimizer that replaces the learning rate with an exponential moving mean of squared delta instead of the difference between the current and updated dataset (Vani, 2019). It is an expansion of Adagrad that addresses the issue of the diminishing learning rate. According to Yaqub (2020), Adadelta approach modifies weights using just first-order time and incurs little computing expense in comparison to other techniques. There is no manual adjusting or learning with this method. In addition, it is resistant to noisy gradient data, data modalities, hyperparameters, and model design decisions. This enhanced the direction of the sharpest fall indicated by a negative gradient. It also enhances the prior algorithm by providing a history window that takes into account a defined number of past gradients during training. Thus, the problem of the diminishing learning rate is avoided.

III. Materials and methods

The flowchart below represents how the Convolutional Neural Network was being made and data processing.

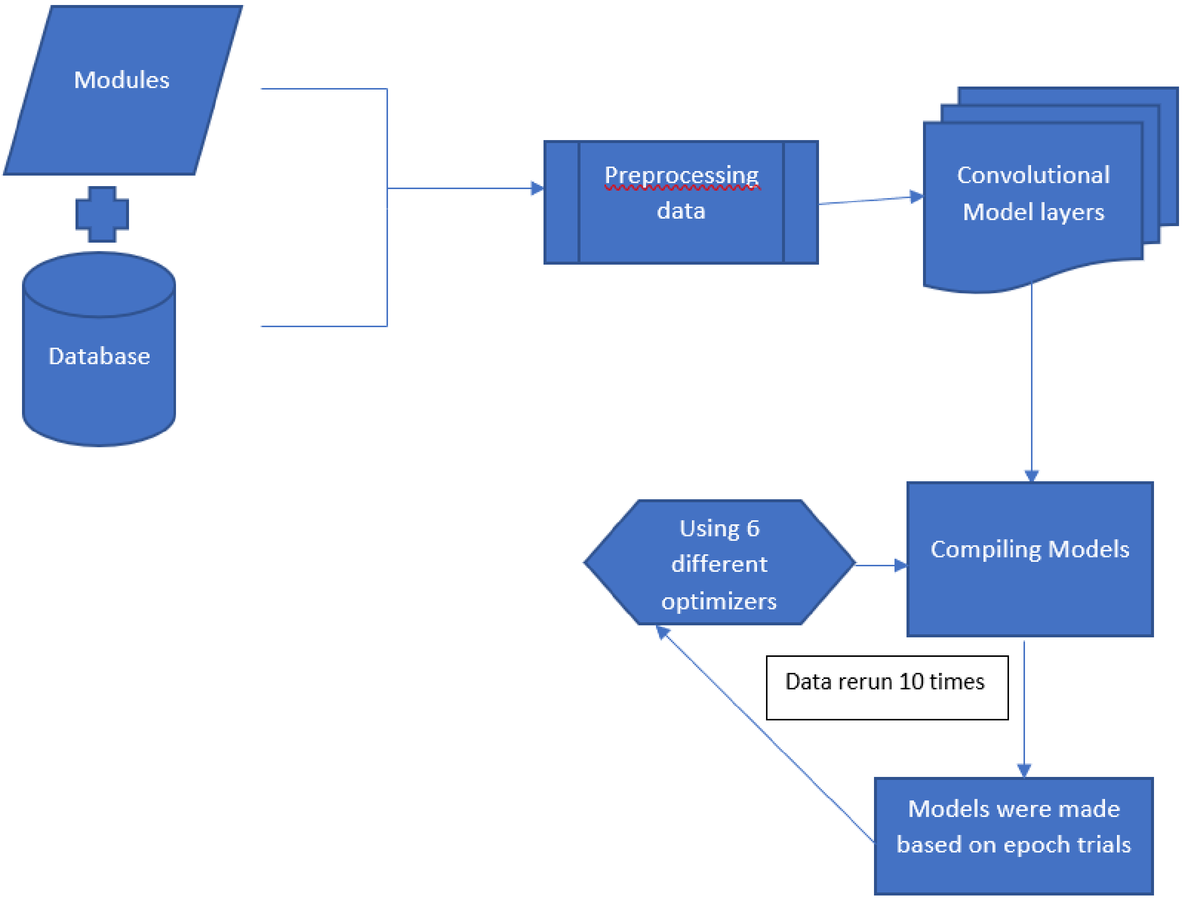

Figure 2. Flowchart of the CNN process

The subchapter that will be divided within the method consists of explanation of the dataset, convolutional neural network model, then the step of preprocessing the data.

A. Schizophrenia Dataset

The dataset used in this project is a public dataset from kaggle. There are two links that was directed to the datasets which are https://www.kaggle.com/broach/button-tone-sz and https://www.kaggle.com/datasets/broach/buttontonesz2. Many researchers used Kaggle to publish their datasets, so it does not count as repository data.

The dataset that was used within the study contains directed tasked patients when they are told to press a button along with auditory tone event-related potentials from 81 patient subjects. The subjects consist of 32 controls with 49 schizophrenia patients. The used dataset was also reduced from the original amount of patients to 15 controls and 15 schizophrenia patients. This is due to the limitations of the applications used to analyze the dataset. Even though the dataset is reduced, a suitable optimizer with a high accuracy value for the CNN model could still be found when analyzing the data.

According to Ford, J. M., et al. (2014), the experiment protocol for the database extraction from the 81 patients began with clinical ratings from the psychiatrist, followed by patient-required tasks. The clinical ratings process began two weeks after the event-related potentials (ERP) assessment, when the pyschiatrist rated the patient's schizophrenia symptoms using the Positive and Negative Syndrome Scale (PANSS), the Scale for Positive Symptoms (SAPS), and the Scale for Negative Symptoms (SNS) (SANS). The patients had to complete two sessions of tasks: the first was a button tone session, and the second was a button alone session.

Before beginning the assignment, patients were instructed to wore an EEG cap, which would be recorded during the sessions (Ford, 2014). The BioSemi ActiveTwo EEG cap with a 64-channel system and 8 external sites was employed. The job began with a button tone session in which patients were required to push a button every 1-2 seconds to provide 1000 Hz tones with a sound pressure level of 80 dB (decibels) and zero latency between button press and tone onset. After delivering 100 tons, the project was terminated. The reserved temporal sequence of tones was stored in the play tone session.

During the button alone session, patients were required to push a button at roughly the same rate while no sound was produced. The EEG signal was continuously digitized with a sampling filter of 1024 Hz and digitally bandpass filtered between 0.5 and 15 Hz during the recording period. The data was then divided into 3000-ms epochs time-locked to button pushes (coincident with the commencement of the tone) and baseline adjusted at -600 to -800 ms (Ford, 2014).

According to figure 1, the modules that were imported using the Google Collab programming language were os, time, numpy, pandas, tqdm, and matplotlib with the pyplot package. Importing sklearn.model train_test_split, to split the trials into train test and split test. Model_selection is a method for creating a blueprint for analyzing data and then applying that blueprint to new data. Choosing the right model for a prediction allows users to get accurate results. normalize is imported by sklearn.preprocessing to normalize the matrix array. In addition, confusion_matrix is imported by sklearn.metrics. Based from the combined dataset, the total number of patients was reduced from 81 participants to 30 participants which consists of 15 healthy patients and 15 schizophrenia patients so that the amount of the participants are balanced between control and schizo patients.

The electrodes that have been used when attached to the patient’s scalp were using 64 EEG channel in order to record their brainwave signals are Fp1, AF7, AF3, F1, F3, F5, F7, FT7, FC5, FC3, FC1, C1, C3, C5, T7, TP7, CP5, CP3, CP1, P1, P3, P5, P7, P9, PO7, PO3, O1, Iz, Oz, POz, Pz, CPz, Fpz, Fp2, AF8, AF4, AFz, Fz, F2, F4, F6, F8, FT8, FC6, FC4, FC2, FCz, Cz, C2, C4, C6, T8, TP8, CP6, CP4, CP2, P2, P4, P6, P8, P10, PO8, PO4, O2, VEOa, VEOb, HEOL, HEOR, TP10, TP10.

Following the input of each module and dataset, as well as the list of electrodes, data processing begins, beginning with the function making, which averages n rows in a matrix. Each path was given one of the two datasets. Before the trials were made, from healthy and schizophrenia patient dataset, it was divided again into two different datasets, which are training dataset and testing dataset, resulted in total of four datasets. Training dataset is a collection of samples used to fit the parameters throughout the learning process while, testing data set, on the other hand, is a set of data that is unrelated to the training data set but has the same probability distribution as the training data set (Dobbin, 2011).

The trials were determined by X and Y counter variables to determine their sets of trials between the test set and the train set, with each having its own set of trials.

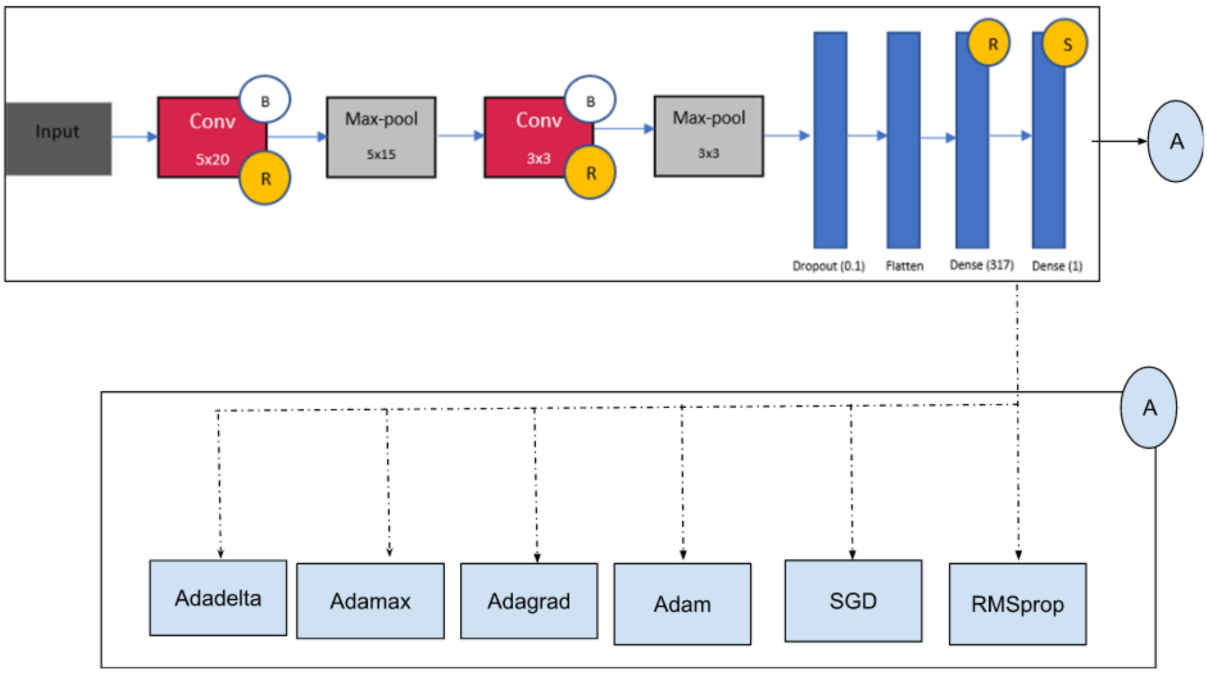

Figure 2. The Proposed Convolutional Neural Network Architecture

B. Convolutional neural network architecture explanation

Within figure 2, many layers are presented within the architecture. Those layers have their own functions and roles within the model, which are:

1. Convolutional layer

A convolutional layer is a filter that uses a weighted matrix to convolve the input. The convolution filter calculates how closely a patch of data resembles a feature. A vertical edge, an arch, or any form can be used as a feature. The weights in the filter matrix are calculated while the data is being trained. Larger filters reflect more global, high-level, and representative information, whereas smaller filters capture as much local information as feasible (Humaidi, 2021).

2. Batch normalization layer

In most deep neural network architectures, the normalized input becomes too big or too small after passing through various adjustments in intermediate layers, causing an internal covariate shift problem that impacts learning (Zhang et al., 2021). To solve this, a batch normalization layer to standardize (mean centering and variance scaling) the input given to the later layers.

3. Max pooling layer

The same parameters apply to pooling layers as they do to convolution layers. Among all the pooling options, Max-Pooling is the most popular. The goal is to reduce the dimensionality of an input representation (image, hidden-layer output matrix, etc.) by retaining the maximum value (active features) in the sub-regions binned (Alzubaidi et al., 2021).

4. Dropout layer

Dropouts are a regularization technique used to prevent model overfitting. Dropouts are utilized to randomly flip a portion of the network's neurons. When these neurons are switched off, both the incoming and outgoing connections to them are likewise severed (Zhang et al., 2021). This is done to facilitate the model's learning.

5. Flatten layer

In essence, flattening involves taking a matrix produced by convolutional and pooling procedures and turning data into a one-dimensional array, according to Humaidi et al. (2021). This is significant since a one-dimensional array makes up the completely linked layer's input.

6. Dense layer

A dense layer, also known as a fully connected layer, is a layer utilized in the neural network's final phases. This layer aids in adjusting the dimensionality of the output from the previous layer so that the model can readily establish the relationship between the values of the data it is working with (Alzubaidi et al., 2021).

7. Activation function layers

Within figure 2, the activation functions that are used in here are relu and sigmoid. The activation layer itself is used to identify whether a neural network's output is yes or no. It converts the values from 0 to 1, -1 to 1, and so on (depending upon the function).

The two activation function layers which are Sigmoid, and ReLU. Sigmoid or logistic activation function is used primarily because it occurs between (0 to 1). As a result, it is particularly useful for models that require us to anticipate the probability as an output (Alzubaidi et al., 2021). The function can be differentiated. The slope of the sigmoid curve between any two locations can be calculated. The function itself is monotonic, but its derivative is not. The logistic sigmoid function has the potential to cause a neural network to become stuck during training. The softmax function is a broader logistic activation function used for multiclass classification.

While for ReLU (Rectified Linear Unit) activation function, it is the most widely utilized activation function in the world. Since then, it has been employed in practically all convolutional neural networks and deep learning algorithms. Both the function and its derivative have a monotonic behavior. However, any negative values are automatically converted to zero, reducing the model's capacity to correctly fit or train from the data. That is, every negative input supplied to the ReLU activation function immediately changes the value to zero in the graph, affecting the output graph by not mapping the negative values effectively (Alzubaidi et al., 2021).

C. Data preprocessing

Following the calculation of the total trials, the architecture model can be built by first determining the layers. Figure 2 illustrates how the architecture is expected to look. In total, there were eight layers. The model mode was set to sequential, and each layer was added one by one, beginning with the first batch of convolutional two-dimensional layers, with kernel_size of five columns and twenty rows, activation layer of relu, and input shape of X_train_2d.shape. A layer of batch normalization was added. With a pool_size of 5 columns and 15 rows, a maximum pooling two-dimensional layer can be created. The next set of layers is a two-dimensional convolutional layer with three columns and three rows of kernel size. A batch normalization layer was also added in this