ResNet50 and GRU: A Synergistic Model for Accurate Facial Emotion Recognition

Shanimol. A[1], J Charles[2]
Research Scholar, Department of Computer Applications Engineering,
Noorul Islam Centre for Higher Education, Tamil Nadu, India[1]
Associate Professor, Department of Computer Applications Engineering,
Noorul Islam Centre for Higher Education, Tamil Nadu, India[2]

***Abstract*—Humans use voice, gestures, and emotions to communicate with one another. It improves oral communication effectiveness and facilitates concept of understanding. Majority of people are able to identify facial emotions with ease, regardless of gender, nationality, culture, or ethnicity. The recognition of facial expressions is becoming more and more significant in a variety of newly developed computing applications. Facial expression detection is a hot topic in almost every industry, including marketing, artificial intelligence, gaming, and healthcare. This study proposes a novel hybrid model combining ResNet-50 and Gated Recurrent Unit (GRU) for enhanced Facial emotion recognition (FER) accuracy. The dataset for the study is taken from Kaggle repository. ResNet-50, a deep convolutional neural network, excels in feature extraction by capturing intricate spatial hierarchies in facial images. GRU, effectively processes sequential data, capturing temporal dependencies crucial for emotion recognition. The integration of ResNet-50 and GRU leverages the strengths of both architectures, enabling robust and accurate emotion detection. Experimental result on CK+ dataset demonstrate that the proposed hybrid model outperforms current methods, achieving a remarkable accuracy of 95.56%. This superior performance underscores the model's potential for real-world applications in diverse domains such as security, healthcare, and interactive systems.**

***Keywords*—Deep convolutional neural network; ResNet-50; Facial Emotion Recognition; Gated Recurrent Unit**

I. INTRODUCTION

Emotions play a major role during communication. A person's mental condition is one of the most important things that can reveal their facial expression. Humans are able to communicate nearly forty five percent of their information verbally and about fifty-five percent nonverbally [1]. Psychologist has defined seven basic emotions such as Disgust, Fear, Surprise Angry, Neutral, Unhappy and Happy [2]. In a broader sense, there are three emotional states of the person. First, neutral emotions, second, positive emotion comprising Happy and Surprise expressions and third, negative emotions comprising Fear, Disgust, Angry and Unhappy expressions. These are basic expressions and are independent of gender and ethnicity [3]. Humans also recognizes other emotions such as contempt, confusion, excitement, stress and desire. Darwin suggested that emotional facial expressions have evolved for a reason.

Nonverbal communication heavily relies on the understanding of facial emotions. Right now automatic face expression recognition is the hardest task thus, there is a strong need for systems that can recognize the same in many different sectors. FER has several uses outside of analyzing behavior and keep track on emotions and mental health of people. It has applications in a variety of domains, including data-driven animation, medical diagnostics, human–robot communication, human–computer interfaces [4], education, robotics, entertainment, holography, smart healthcare systems, security systems, criminology [5], and stress detection [6–7]. Facial expressions are becoming increasingly significant in the medical sciences, especially for bipolar patients whose mood swings are frequent.

FER is also beneficial for applications like smart card readers, social robots, e-learning, criminal justice systems, and customer satisfaction identification [8-9]. The classic emotion identification system consists of three key blocks namely feature extraction, face detection, and emotion classification. Conventional methods of emotion detection have the disadvantage of mutually optimizing feature extraction and categorization. The automation of face emotion recognition and classification is a difficult task. A few fundamental emotions are used by the research community, including fear, anger, upset, and pleasure. However, machines find it extremely difficult to distinguish between a wide ranges of emotions. The major contributions of the proposed research work are follows:

- Developed an efficient Facial Emotion Recognition (FER) method utilizing a deep learning model.
- Implemented a hybrid deep learning approach by integrating ResNet 50 with Gated Recurrent Unit (GRU).
- Achieved enhanced accuracy in Facial Emotion Recognition.

The rest of the paper is organized as follows: In Section II, a summary of literature is provided, highlighting areas that indicate a need for more investigation. In Section III, the methodology is explained in depth. Section IV goes into great detail about the results that the suggested strategy produced. A discussion is provided in Section V and finally, a summary of the findings is included in Section VI, which gives a conclusion to the paper.

II. Literature Review

A framework that used a BiLSTM fusion network and simultaneously learned temporal dynamics and spatial information for FER was presented by Liang et al. [10]. Three benchmark databases—namely MMI, CK+, Oulu-CASIA, were used in the experiment. The technique learned discriminative spatial features and short-term dynamic features using two separate CNNs, and then combined them at the feature level. A comparison of the model's performance using the Oulu-CASIA dataset revealed an accuracy of 91.07%. Because there were only few training samples available for the study, the method's generalizability was constrained. A spatio-temporal feature representation learning method for FER that was resistant to changes in expression intensity was presented by Kim et al. [11]. Regardless of the intensity of the expression, the approach made use of representative expression described in face sequences. Using a CNN, spatial properties were learned. Long short-term memory of the face expression was used to learn the temporal property of the spatial feature representation. The studies were carried out using two datasets namely one for spontaneous micro-expression (CASME II) and the other for purposeful expression (MMI). The accuracy of the approach was found to be 72.83% and 78.61% in the intra- and inter-dataset evaluations.

A Transfer learning approach to emotion recognition was proposed by Chowdary et al. [12]. Pre-trained vgg19, Resnet50, and Mobile Net, Inception V3, networks were used in this work. The CK + database was used in the experiment and 94.2% accuracy was attained with the help of MobileNet. Using a CNN, Debnath et al.'s new facial emotional recognition model [13] identified seven distinct emotions from image data. The model attained an accuracy of 92.05% using the generalization strategy on the JAFFE dataset. A modular framework was presented by Alreshidi et al. [14] for the classification of facial emotions into seven distinct states. The authors failed to utilize geometric elements that could enhance the performance. The approach achieved 59.0% accuracy and might be used to treat and diagnose patients with emotional problems.

A Custom CNN Architecture was presented by Borgalli et al. [15] to do fundamental FER in static images. The three datasets utilized in the methodology to evaluate the model are FER13, JAFFE, and CK+. The CK+ datasets achieved an accuracy rate of 92.27% on fundamental emotions. The method's significant recognition error amount was one of its limitations. In order to circumvent the conventional feature extraction procedure, Bukhari et al. [16] created a CNN model that was utilized as a feature extractor for emotion identification using facial expression. Three pre-trained models were employed in this work by the authors: VGG-16, ResNet-50, and Inception-V3. The accuracy rates for CNN 92.91% on ck+ dataset, according to the experimental results.

CNN, which predict and assign probabilities to each emotion, were the basis of an efficient facial emotion identification system for the seven basic human emotions presented by Ghaffar et al. [17]. In order to improve prediction, the system applied a variety of preprocessing procedures to each image as deep learning models learn from data. To include each image in the training dataset, the face detection algorithm was run on each one first. With the combined dataset, the method's maximum accuracy was 78.1%. In order to overcome the facial expression recognition (FER) problem, Kandhro et al. [18] proposed a CNN; in recent years, more and more significant efforts have been done in this area. This FER technique can be used to obtain facial expressions based on regularization settings, activations, and optimizations from databases such as CK+ and JAFFE. Numerous techniques, such as regularization, optimization, and activation, in addition to additional hyperparameters, were used to assess the model's performance. The authors obtained 71% test accuracy and 97% training accuracy using the FER2013 dataset. By utilizing several facial features with appropriate dimensions space reduction and applying a kernel filter as part of the preprocessing technique, Kumar Arora et al. [19] suggested the facial feature for emotional recognition using a deep learning algorithm.

A. Research Gap

Facial expression recognition heavily depends on facial landmarks in image-based approaches, which can be prone to errors in varying conditions. Model-based approaches, while accurate, rely on intense numerical computations due to the need for complex mapping functions, making them resource-intensive and time-consuming when training on large datasets. Current models exhibit good performance but require further improvement to handle real-world variations such as different lighting conditions, occlusions, and diverse facial expressions across various ethnicities and age groups. Existing methods often fall short in robustness and generalization, partly due to training on relatively small or biased datasets. Additionally, many FER systems are not optimized for real-time operation on edge devices with limited computational resources. Fourier transform techniques, commonly used in these systems, may miss important spatial features crucial for emotion detection as they focus primarily on frequency domain information. Addressing these gaps is essential for developing FER systems that are both efficient and applicable in diverse, real-world scenarios.

III. Materials and Methods

The proposed methodology as shown in Fig. 1 comprises two main components: a pre-trained ResNet50 [20] for image feature extraction and a GRU [21] for capturing sequential patterns. ResNet-50 was chosen for its exceptional feature extraction capabilities, which are crucial for capturing the intricate spatial details of facial expressions, while GRU was integrated to leverage its strength in sequential pattern recognition, enabling the model to effectively analyze the temporal dynamics of emotions. Existing methods, such as those relying solely on Convolutional Neural Networks (CNNs) or Long Short-Term Memory (LSTM) networks, often fall short in either spatial feature extraction or temporal sequence modeling, making them less effective for FER tasks that require a holistic approach. The proposed method overcomes these limitations by combining the strengths of both architectures, making it more suitable for the complex pattern recognition required in accurately classifying emotions.

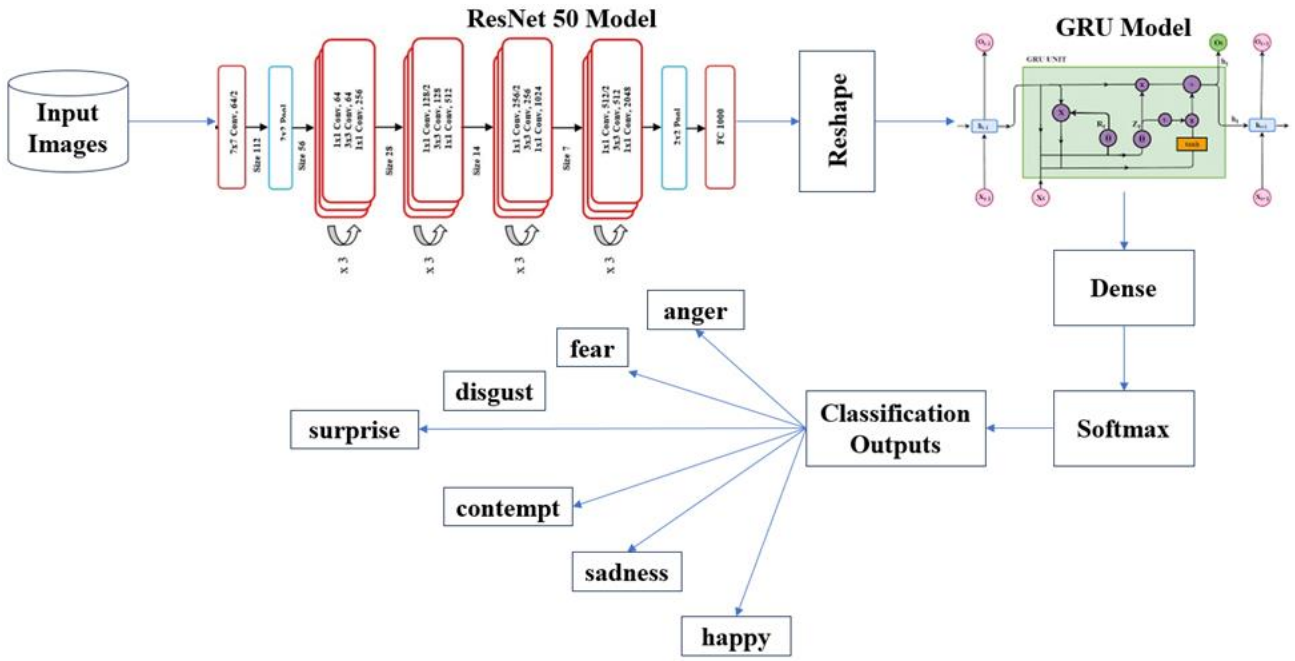

Fig. 1. Block diagram of the proposed model.

Following dataset collection, pre-processing and augmentation techniques are applied. The data generator resizes images to a target size of 224x224 pixels, rescale pixel value to the range [0, 1], and apply shear transformations, zoom in/out, horizontal and vertical flips, and rotations up to 30 degrees. The images are batched into groups of 32, and class labels are encoded in categorical format. This setup is crucial for training and evaluating the model on the CK+ dataset, enhancing generalization through data augmentation. The ResNet50 and GRU networks are integrated into a hybrid model by concatenating their outputs. The ResNet and GRU components are connected sequentially, with the GRU input reshaped to match its expected input shape. The performance of the model is evaluated on the several performance measures.

A. *Dataset Description*

Data is sourced from Kaggle repository https://www.kaggle.com/datasets/davilsena/ckdataset/. The sample images are shown in Fig. 2. Dataset Contains modified data up to 920 images from 920 original CK+ dataset. Data is already reshaped to 48x48 pixels, in grayscale format and face cropped using haarcascade_frontalface_default. Noisy images were adapted to be clearly identified using Haar classifier.

B. *Data Pre-processing and Augmentation*

The preprocessing and data augmentation steps for the model involve several techniques to enhance training and evaluation on the CK+ dataset. Initially, the dataset is preprocessed by rescaling pixel values to the range [0, 1]. The data generator is then configured to apply various augmentation techniques, including shear transformations, zooming in and out, horizontal and vertical flips, and rotations up to 30 degrees. Images are resized to a target size of 224x224 pixels and batched into groups of 32. Additionally, class labels are encoded in categorical format. This comprehensive setup is crucial for improving the model's generalization capabilities through effective data augmentation.

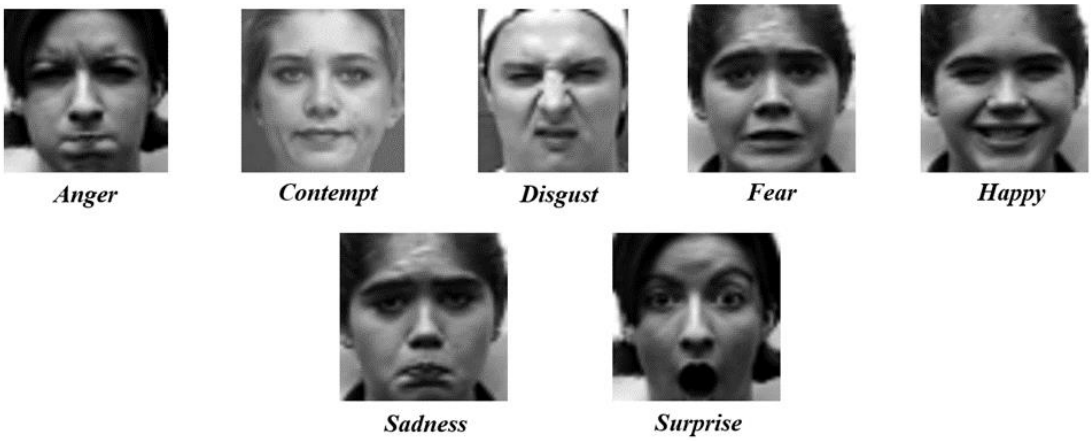

Fig. 2. Sample images from the dataset.

C. ResNet 50 Architecture

Four main parts of the ResNet50 architecture are the identity block, fully connected layer, convolutional block, and convolutional layer. Fig. 3 shows the architecture of the ResNet 50 model. The features that the convolutional layers have extracted from the input image are being processed and transformed by the identity block and convolutional block. The identity block is a straightforward block that adds the input back to the output after passing it through several convolutional layers. The network is able to learn residual functions, which convert input into desired output. In the convolutional layers of ResNet50, batch normalization and ReLU activation come after many convolutional layers. These layers are in charge of taking characteristics like edges, textures, and forms out of the input image. Max pooling layers, which minimize the spatial dimensions of the feature maps while maintaining the most crucial properties, come after convolutional layers. The fully connected layers make up the last section of ResNet50. The last classification is determined by these layers. The output of the final fully connected layer is fed into a softmax activation function to get the final class probabilities.

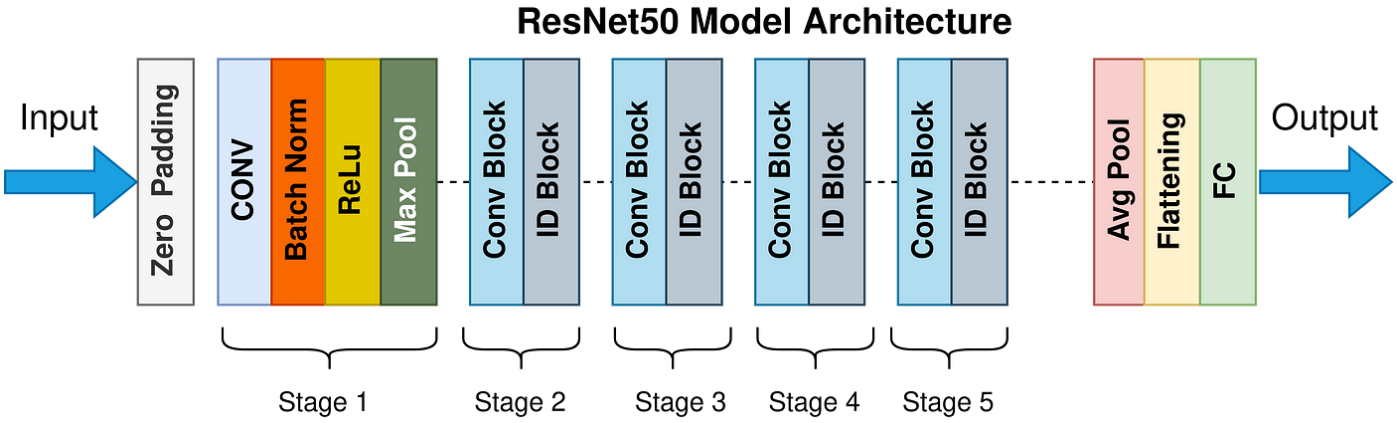

Fig. 3. Architecture of ResNet50 model.

When deep neural networks are trained, an issue known as "vanishing gradients" might arise. This is when the parameter gradients in the deeper layer get very small, which makes it harder for such layers to learn. The deeper the network, the more severe this issue gets. By enabling data to move straight from the network's input to its output and omitting one or more tiers, skip connections solve this issue. Instead of having to learn the complete mapping from scratch, the network can learn residual functions that convert input into the intended output. The residual block is depicted in Fig. 4. The output of the residual block is represented by Eq. (1).

$$Y = F(X, \{W_m\} + X) \quad (1)$$

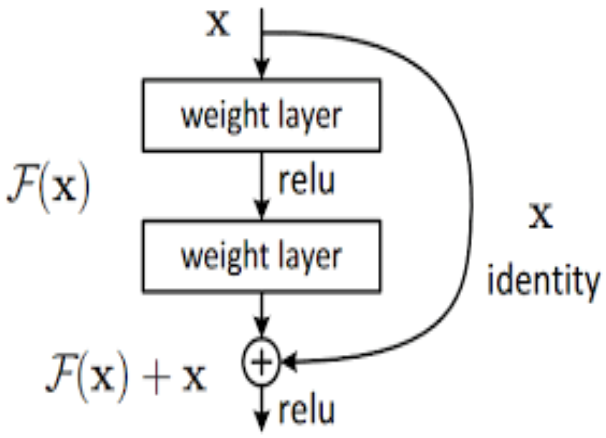

Fig. 4. Block diagram of the residual block.

The input to the residual block is denoted by X, Y is the output of the block, W_m represents the weights of the convolutional layers within the block, and F is the residual function. In a residual block, the residual function F typically consists of two or three convolutional layers. For a three-layer residual block, the residual function can be expressed as in Eq. (2).

$$F(X, \{W_m\} = W3\sigma(W2\sigma(W1X))F(X, \{Wm\}) = W3\sigma(W2\sigma(W1X)) \quad (2)$$

σ denotes the ReLU activation function, W1,W2,W3 are the weights of the convolutional layers. The identity mapping X sometimes be transformed using a linear projection to match the dimensions of $F(X, W_m)$ when necessary. This can be done using a 1x1 convolution as depicted in Eq. (3).

$$Y = F(X, \{W_m\}) + W_s X \quad (3)$$

where W_s is the weight matrix of the 1x1 convolution used for matching dimensions. The activation function used is typically the Rectified Linear Unit (ReLU), expressed as in Eq. (4).

$$\sigma(x) = max(0, x) \quad (4)$$

Batch normalization as expressed in Eq. (5) is applied after each convolution and before the activation function:

$$BN(x) = \frac{x-\mu}{\sigma^2+\epsilon} * \gamma + \beta \quad (5)$$

Where μ and σ2 are the batch mean and variance, γ and β are learnable parameters, and ϵ is a small constant to prevent division by zero.

D. Gated Recurrent Unit

By allowing information to be selectively retained or lost over time, GRU is intended to mimic sequential data. Because GRU has fewer parameters and a simpler architecture, it may be easier to train and use less computing power. The update gate and the reset gate are the two distinct gates that are part of the GRU architecture as shown in Fig. 5. The distinct functions of each gate greatly add to the high efficiency of the GRU. Long-term connections are recognized by the update gate, whereas short-term ties are identified by the reset gate.

The computation steps of a GRU in the Reset Gate, update gate, candidate hidden state, as are the following.

$$R_t = \sigma(A_{x,z}X_t + A_{H,z}H_{t-1} + B_z) \quad (6)$$

$$Z_t = \sigma(A_{x,z}X_t + A_{H,z}H_{t-1} + B_z) \quad (7)$$

$$\acute{H}_t = \tanh(A_{H,H}R_tH_{t-1}) + A_{x,H}X_t + B_H \quad (8)$$

$$H_t = (1 - Z_t)H_{t-1} + Z_t\acute{H}_t \quad (9)$$

where, $\acute{H}_t$ is the candidate hidden state that is incorporated proportionately to the hidden state, R_t is the reset gate value, and Z_t is the update gate.

E. Proposed ResNet 50-GRU Hybrid Model

The proposed hybrid model for FER integrates a pre-trained ResNet-50 and a GRU. The model architecture of the proposed hybrid model is depicted in Fig. 6. The model takes an image input of shape (224, 224, 3), with the ResNet and GRU components connected sequentially, reshaping the GRU input to match its expected input shape. The ResNet-50, configured as a feature extractor by removing its top layers and applying global average pooling, reduces spatial dimensions and is followed by a dense layer and ReLU activation to enhance feature representation. Concurrently, a GRU network with 128 units is constructed, also followed by a dense layer with 256 units and ReLU activation. These two networks are then integrated by concatenating their outputs, and the final output layer is a dense layer with softmax activation, which produces probability distributions over seven emotion classes.

F. Hardware and Software Setup

The computational setup for this research utilized a machine with robust specifications, featuring an Intel Core i7 processor. 32GB of RAM, and the formidable NVIDIA GeForce GTX 1080Ti GPU. Model implementation was seamlessly carried out through the Keras library, functioning as a prototype built upon the Tensorflow framework and executed using the versatile Python language. Keras, known for its user-friendly interface and powerful capabilities, proved instrumental in crafting intricate Neural Network architectures. This framework ensures efficient utilization of computing resources, seamlessly accommodating CPU, GPU, and TPU environments. To leverage extensive computational capabilities and streamline model training, the deployment was orchestrated on Google Colab. This cloud-based Python notebook environment not only provides complimentary access to robust computational resources but also facilitates collaborative development, making it an optimal choice for training models.

Hyper parameters are essential configuration settings that define the behaviour and characteristics of a machine learning framework throughout the training process. Unlike the parameters of the model, which are learned from the data itself, hyper parameters are set by the user before training begins. The neural network model uses the Adam optimizer. The training process is guided by the Categorical cross-entropy loss function. During training, the model processes input data in batches of 32 samples per iteration. The training is carried out over 50 epochs, signifying the number of times the entire training dataset is processed by the model. These hyper parameter choices, such as the optimizer, loss function, batch size, and number of epochs, collectively define the configuration for training the neural network model, aiming to optimize its performance on the proposed emotion detection. The model configuration of the suggested approach is tabulated in Table I.

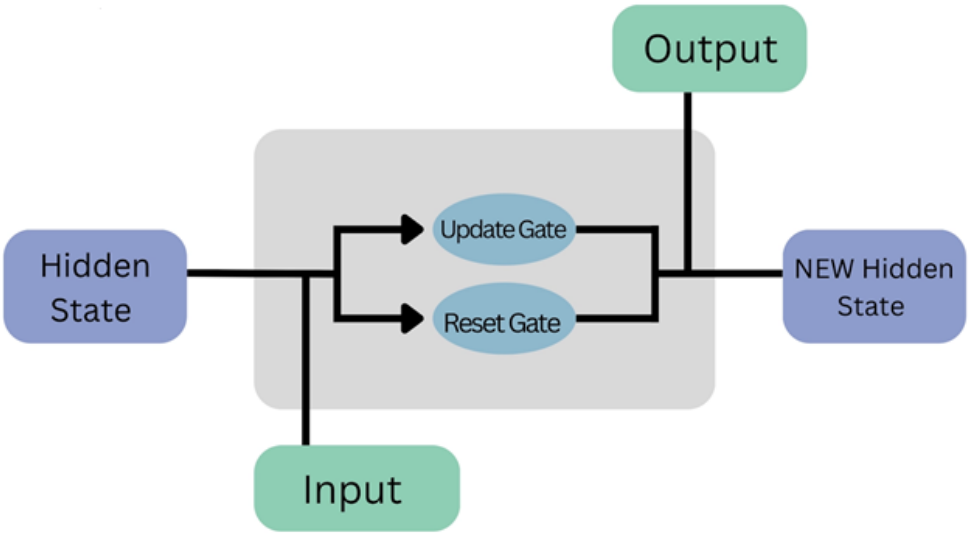

Fig. 5. Architecture of Gated Recurrent Unit.

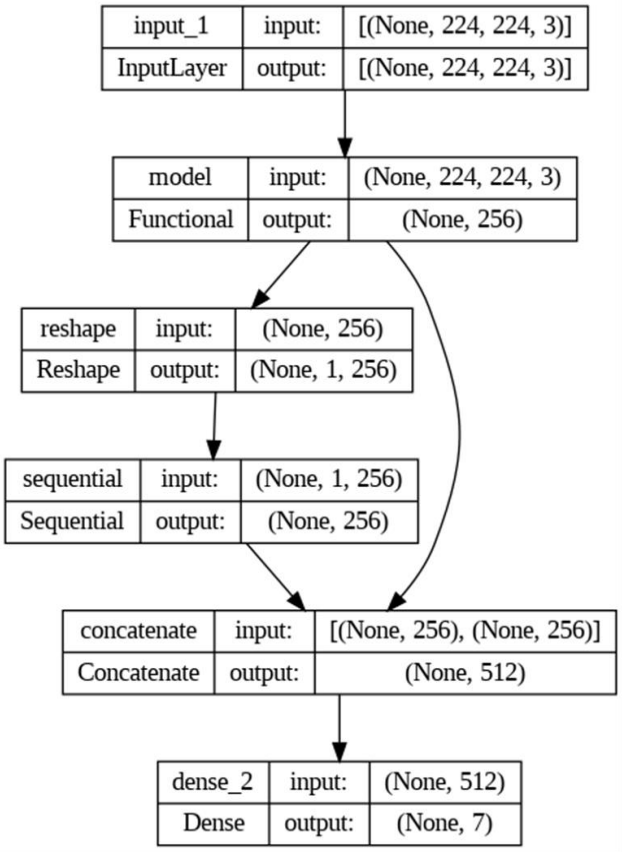

Fig. 6. Proposed model architecture.

TABLE I. MODEL CONFIGURATIONS

Hyper parameter	Values
Optimizer	Adam
Loss Function	Categorical crossentropy
No. of epochs	50
Batch Size	32
Activation Function	ReLu, Softmax

IV. EXPERIMENTAL RESULTS

The accuracy and loss plots are essential tools for assessing the performance and learning dynamics of the proposed emotion classification model. The accuracy plot visually depicts how accurately the model predicts the emotional labels of the data across training iterations for both the training and validation datasets. This plot tracks the consistency between the model's predictions and the actual emotional labels, serving as a crucial indicator of the model's performance throughout the training process. The accuracy plot highlights the model's ability to effectively distinguish between different facial emotions during training. Ideally, in the early epochs, both training and validation accuracies rise simultaneously, demonstrating the model's ability to generalize its knowledge beyond the training dataset.

The accuracy plot of the model is shown in Fig. 7. In the initial epochs of training, the proposed system demonstrates high accuracy, starting at 98.74% in epoch 1. This strong initial performance indicates the model's rapid learning capacity. As the training progresses, accuracy consistently remains high, reaching 99.20% by epoch 2 and continuing to show incremental improvements. However, some fluctuations in accuracy occur between epochs 13 and 16, where accuracy drops from 98.40% to 94.04%, reflecting a temporary period of instability. Despite these fluctuations, the model quickly recovers, achieving a perfect accuracy of 100% by epoch 10 and maintaining this level through the final epochs, from epoch 48 to epoch 50. These fluctuations are typical in deep learning training processes, often due to batch variance or learning rate adjustments.

Fig. 7. Accuracy plot of the proposed system.

The difference between the predicted emotions and the true labels is quantified as the model's loss, which is illustrated in the loss plot. Throughout the training process, the objective is for the loss to decrease steadily, reflecting that the model is improving its predictions and reducing errors with each iteration, as depicted in Fig. 8.

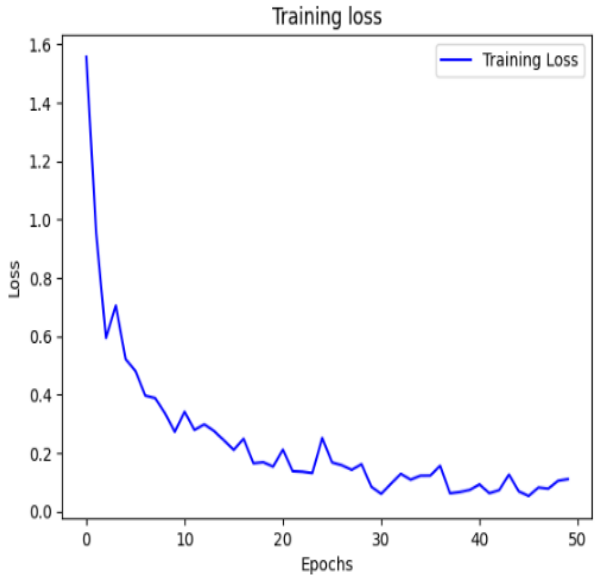

Fig. 8. Loss plot of the proposed system.

In the initial epochs, the loss decreases steadily from 0.0329 to 0.0010, and the accuracy increases from 0.9874 to 1.0000. By epoch 22, the loss decreases to 0.0911 with the same accuracy, but in epoch 23, the loss increases to 0.1567, and accuracy slightly drops to 95.53%, showing some initial fluctuation. In the final epochs, the loss significantly decreases, with epoch 48 showing a loss of 0.0021 and 100% accuracy, continuing to epoch 49 with a loss of 0.0011 and epoch 50 reaching the lowest loss of 0.0010, both maintaining 100% accuracy. This indicates that the model is learning well and improving its performance over time. However, there are fluctuations in the loss and accuracy throughout the epochs, such as a slight increase in loss during epochs 11 and 12, followed by a decrease. Overall, the fluctuations seem relatively minor, and the model achieves a very low loss and high accuracy by the final epoch, suggesting that it has effectively learned the patterns in the data and generalized well.

A valuable method for assessing the effectiveness of the proposed emotion classification system is the use of a confusion matrix. This matrix offers a structured overview of the model's performance by comparing its predicted emotional categories with the actual labels across various classes. It organizes the outcomes into a table format, where the rows correspond to the true emotional labels and the columns correspond to the predicted labels, as illustrated in Fig. 9. Each cell within the matrix displays the count of instances where the model's predictions either match or deviate from the true emotional labels. The confusion matrix is divided into four quadrants, with the diagonal elements representing correct predictions and the off-diagonal elements indicating misclassifications.

Performance metrics derived from the confusion matrix offer a thorough evaluation of the proposed model's efficacy in classifying emotions. The performance of the system is mainly evaluated on four parameters accuracy, precision, recall, F1-score. These measures, which are based on the concepts of False Positive (FP), False Negative (FN), True Negative (TN), and True Positive (TP), are essential for assessing the model's performance. The calculation of accuracy involves dividing the total number of predictions by the number of right predictions.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

The exactness of a prediction is measured by its precision, or the number of true positives. Instead, recall quantifies completeness, or the number of real positives that were anticipated as positives.

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

$$F1-Score = 2 * \left(\frac{Precision*Recall}{Precision+Recall}\right) \quad (13)$$

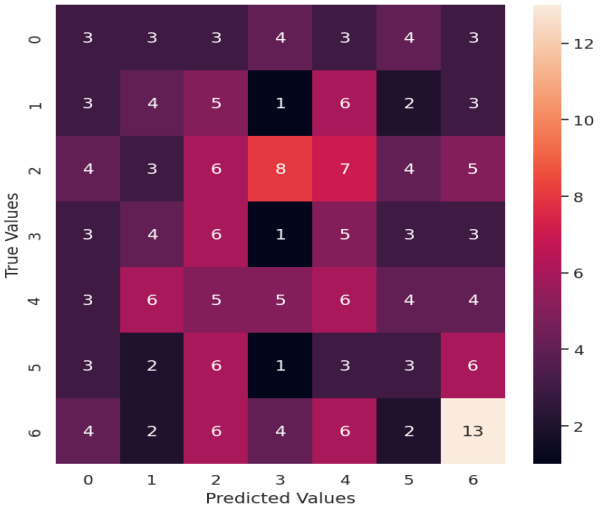

Fig. 9. Confusion matrix.

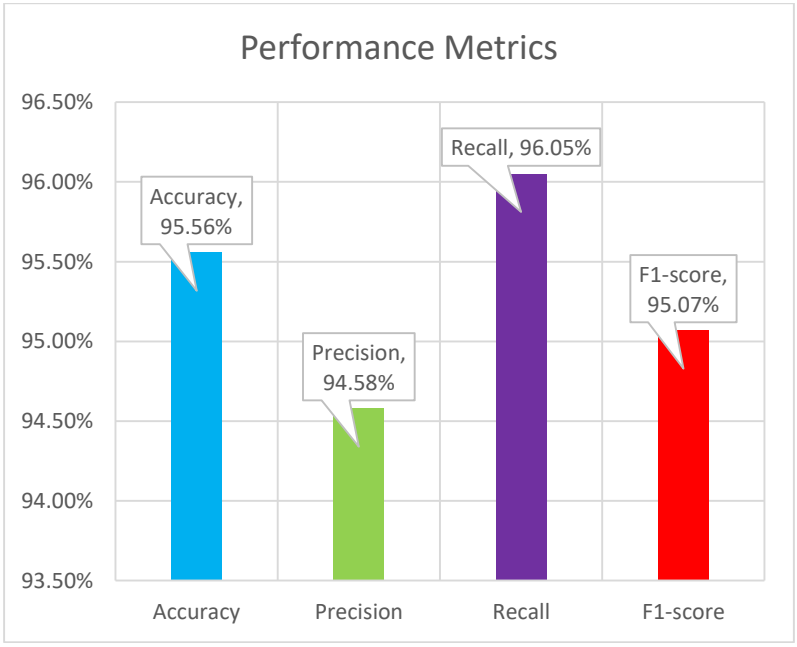

Fig. 10. Performance metrics.

The performance metrics of the proposed emotion classification system are detailed in Fig. 10, illustrating its effectiveness in accurately classifying emotions. The accuracy metric stands at 95.56%, indicating the overall correctness of the model's predictions compared to the total number of predictions made. Precision, measured at 94.58%, reflects the model's capability to correctly identify specific emotions among those predicted. Recall, which measures at 96.05%, signifies the model's ability to accurately retrieve all instances of a particular emotion from the dataset. The F1-score, calculated at 95.07%, harmonizes precision and recall into a single metric, offering a balanced assessment of the model's performance in emotion classification tasks. These metrics collectively demonstrate the system's high accuracy and reliability in recognizing and distinguishing between different emotional states, underscoring its potential for practical applications requiring nuanced emotion detection.

The prediction results of the proposed emotion classification system are illustrated in Fig. 11. This figure provides a comprehensive visual representation of the model's performance, showcasing how effectively it can identify and classify various emotional expressions. By examining the prediction results, one can assess the accuracy and reliability of the model in real-world scenarios. The figure highlights the model's capability to distinguish between different emotions, demonstrating its potential effectiveness and practical application in emotion recognition tasks.

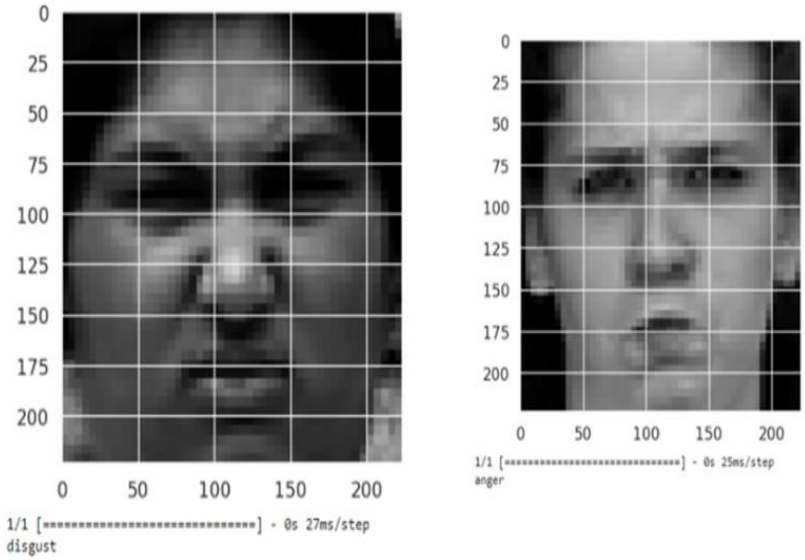

Fig. 11. Prediction output.

V. Discussion

Comparing the performance of the proposed hybrid emotion classification network against existing methods primarily based on machine learning and deep learning is a pivotal aspect of this study. Table II and Fig. 12 provides a comparative analysis that highlights the effectiveness of the hybrid model by carefully evaluating outcomes in contrast to established approaches. This model addresses the challenges of accurately recognizing and classifying emotions from facial expressions, which is critical in fields such as human-computer interaction, healthcare, and security. This evaluation rigorously examines a range of metrics and parameters to assess how well the proposed method performs compared to traditional methodologies used in emotion recognition. The aim is to demonstrate the superiority and robustness of the hybrid approach in accurately classifying emotions, showcasing its potential to outperform conventional techniques in real-world applications.

The comparison report presents various state-of-the-art models and their corresponding methodologies and results in a certain task, likely classification or prediction. Liang et al. employed a Bidirectional Long Short-Term Memory (BiLSTM) model achieving an accuracy of 91.07%. Kim et al. utilized a CNN-LSTM hybrid model, attaining a slightly lower accuracy of 78.61%. Debnath et al. deployed a CNN achieving a higher accuracy of 92.05%, while Borgalli et al. introduced a custom CNN with a slightly higher accuracy of 92.27%. Bukhari et al. implemented a ResNet-50 model, achieving an accuracy of 92.91%. Finally, the proposed model in the report, a hybrid of ResNet-50 and GRU, demonstrated the highest accuracy of 95.56%. This comparison highlights the effectiveness of different architectures and demonstrates the superiority of the proposed hybrid model, which integrates both convolutional and recurrent neural network components, achieving the highest accuracy among the compared methods.

TABLE II. Comparison of Proposed Model with Existing Methods

Authors	Methodology	Result
Liang et al [10]	Bi-LSTM	91.07%
Kim et al [11]	CNN-LSTM Hybrid model	78.61%
Debnath et al [13]	CNN	92.05%
Borgalli et al [15]	Custom CNN	92.27%
Bukhari et al [16]	ResNet -50	92.91%
Proposed model	**Hybrid model ResNet 50 and GRU**	**95.56%**

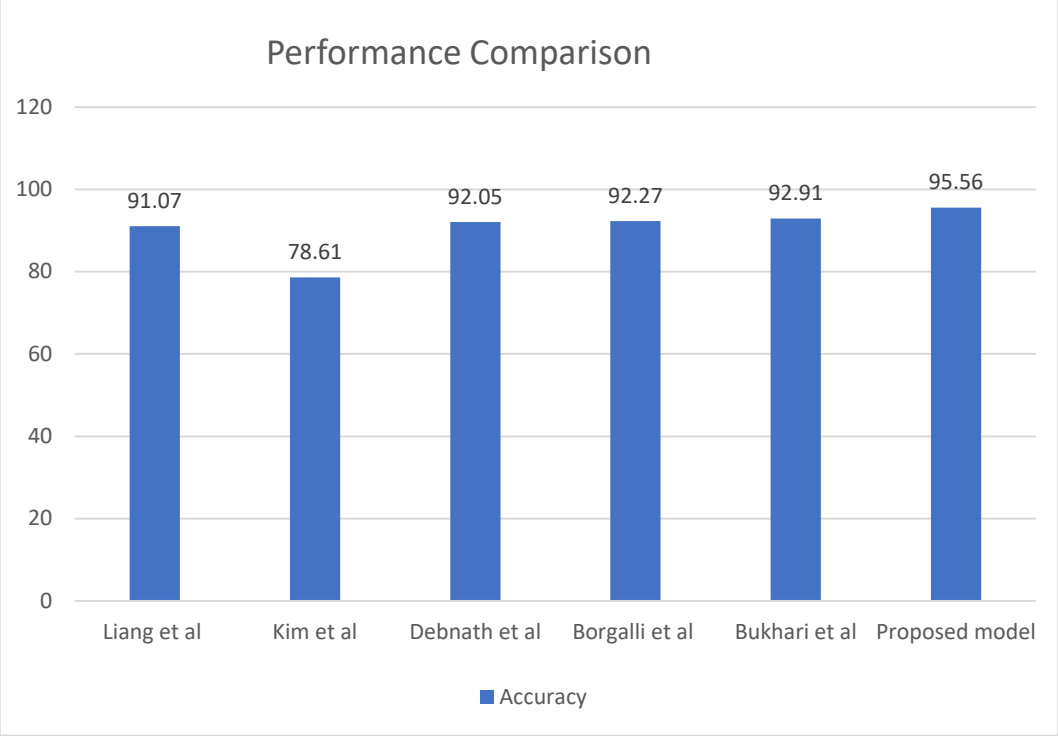

Fig. 12. Performance comparison.

Moreover, the robustness of the proposed model in handling diverse and complex emotional expressions demonstrates its potential for real-world applications where emotion recognition must be accurate and reliable. The comparison study underscores the model's capability to outperform conventional techniques, making it a promising solution for advancing the field of emotion recognition. This hybrid approach not only achieves higher accuracy but also offers a more nuanced understanding of emotional expressions, paving the way for more sophisticated and effective emotion recognition systems.

VI. Conclusion

Facial expressions are a vital tool for determining human emotions since they are reflections of those emotions. The majority of the time, a person's facial expressions are a nonverbal means of expressing their emotions. These emotions can be used as concrete evidence to determine whether or not someone is telling the truth. This study aimed to classify facial expressions into one of seven emotions using the CK+ dataset. The study successfully demonstrates the efficacy of a hybrid model combining a pre-trained ResNet50 and a GRU for FER. By leveraging the powerful feature extraction capabilities of ResNet50 and the sequential pattern recognition strengths of GRU, the proposed model achieves an accuracy of 95.56% in classifying emotions. When compared to cutting-edge outcomes, the developed model provides good accuracy. This hybrid approach not only validates the integration of ResNet 50 and GRU for complex pattern recognition tasks but also sets a robust framework for future research in FER and related fields.

While the proposed model has shown promising results, there are several areas for future work to further enhance the performance and applicability of FER systems. First, expanding the dataset to include more diverse and real-world scenarios could improve the model's generalizability, making it more robust in different environments and cultures. Additionally, exploring the integration of other advanced deep learning architectures, such as Transformer-based models, could potentially enhance the model's ability to capture complex dependencies in facial expressions. Moreover, incorporating multimodal data, such as voice and physiological signals, alongside facial expressions could lead to more comprehensive emotion recognition systems.

References

[1] Salim, M. S. (2023). Verbal and non-verbal communication in linguistics. International Journal of Innovative Technologies in Social Science, (2 (38)).

[2] Frenzel, A. C., Daniels, L., & Burić, I. (2021). Teacher emotions in the classroom and their implications for students. Educational Psychologist, 56(4), 250-264.

[3] Chen, Y., & Joo, J. (2021). Understanding and mitigating annotation bias in facial expression recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14980-14991).

[4] Rajesh Kumar, G., Srinivasa Rao, D., Rajasekhar, N., Ramesh Babu, C., Rohini, C., Ravi, T., & Mangathayaru, N. (2022, November). Emotion Detection Using Machine Learning and Deep Learning. In International Conference on Intelligent Computing and Communication (pp. 705-715). Singapore: Springer Nature Singapore.

[5] Channing, I., Churchill, D., & Yeomans, H. (2023). Renewing historical criminology: Scope, significance, and future directions. Annual Review of Criminology, 6, 339-361.

[6] Mansour, R. F., El Amraoui, A., Nouaouri, I., Díaz, V. G., Gupta, D., & Kumar, S. (2021). Artificial intelligence and internet of things enabled disease diagnosis model for smart healthcare systems. IEEE Access, 9, 45137-45146.

[7] Goel, R., & Gupta, P. (2020). Robotics and industry 4.0. A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development, 157-169.

[8] Hassouneh, A., Mutawa, A. M., & Murugappan, M. (2020). Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. Informatics in Medicine Unlocked, 20, 100372.

[9] Otamendi, F. J., & Sutil Martín, D. L. (2020). The emotional effectiveness of advertisement. Frontiers in Psychology, 11, 563695.

[10] Liang, D., Liang, H., Yu, Z., & Zhang, Y. (2020). Deep convolutional BiLSTM fusion network for facial expression recognition. The Visual Computer, 36, 499-508.

[11] Kim, D. H., Baddar, W. J., Jang, J., & Ro, Y. M. (2017). Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Transactions on Affective Computing, 10(2), 223-236.

[12] Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2023). Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications, 35(32), 23311-23328.

[13] Debnath, T., Reza, M. M., Rahman, A., Beheshti, A., Band, S. S., & Alinejad-Rokny, H. (2022). Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity. Scientific Reports, 12(1), 6991.

[14] Alreshidi, A., & Ullah, M. (2020, February). Facial emotion recognition using hybrid features. In Informatics (Vol. 7, No. 1, p. 6). MDPI.

[15] Borgalli, M. R. A., & Surve, S. (2022, March). Deep learning for facial emotion recognition using custom CNN architecture. In Journal of Physics: Conference Series (Vol. 2236, No. 1, p. 012004). IOP Publishing.

[16] Bukhari, N., Hussain, S., Ayoub, M., Yu, Y., & Khan, A. (2022). Deep learning based framework for emotion recognition using facial expression. Pakistan Journal of Engineering and Technology, 5(3), 51-57.

[17] Ghaffar, F. (2020). Facial emotions recognition using convolutional neural net. arXiv preprint arXiv:2001.01456.

[18] Kandhro, I. A., Uddin, M., Hussain, S., Chaudhery, T. J., Shorfuzzaman, M., Meshref, H., ... & Khalaf, O. I. (2022). Impact of activation, optimization, and regularization methods on the facial expression model using CNN. Computational Intelligence and Neuroscience, 2022.

[19] Kumar Arora, T., Kumar Chaubey, P., Shree Raman, M., Kumar, B., Nagesh, Y., Anjani, P. K., ... & Debtera, B. (2022). Optimal facial feature based emotional recognition using deep learning algorithm. Computational Intelligence and Neuroscience, 2022(1), 8379202.

[20] Koonce, B., & Koonce, B. (2021). ResNet 50. Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization, 63-72.

[21] Dey, R., & Salem, F. M. (2017, August). Gate-variants of gated recurrent unit (GRU) neural networks. In 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS) (pp. 1597-1600). IEEE.