Advancing Human Action Recognition and Medical Image Segmentation using GRU Networks with V-Net Architecture

Dustakar Surendra Rao[1], L. Koteswara Rao[2*], Vipparthi Bhagyaraju[3], P. Rohini[4]
Department of ECE, Koneru Lakshmaiah Education Foundation, Aziz Nagar, Hyderabad, 500075, Telangana, India[1, 2*]
Department of ECE, Guru Nanak Institutions Technical Campus, Hyderabad, 501506, Telangana, India[1]
Department of ECE, Siddhartha Institute of Engineering and Technology, Hyderabad, 501506, Telangana, India[3]
Department of Data Science and Artificial Intelligence, ICFAI Foundation for Higher Education, Hyderabad, India[4]

***Abstract*—Human Action Recognition and Medical Image Segmentation study presents a novel framework that leverages advanced neural network architectures to improve Medical Image Segmentation and Human Action Recognition (HAR). Gated Recurrent Units (GRU) are used in the HAR domain to efficiently capture complex temporal correlations in video sequences, yielding better accuracy, precision, recall, and F1 Score than current models. In computer vision and medical imaging, the current research environment highlights the significance of advanced techniques, especially when addressing problems like computational complexity, resilience, and noise in real-world applications. Improved medical image segmentation and human action recognition (HAR) are of growing interest. While methods such as the V-Net architecture for medical picture segmentation and Spatial Temporal Graph Convolutional Networks (ST-GCNs) for HAR have shown promise, they are constrained by things like processing requirement and noise sensitivity. The suggested methods highlight the necessity of sophisticated neural network topologies and optimisation techniques for medical picture segmentation and HAR, with further study focusing on transfer learning and attention processes. A Python tool has been implemented to perform min-max normalization, utilize GRU for human action recognition, employ V-net for medical image segmentation, and optimize with the Adam optimizer, with performance evaluation metrics integrated for comprehensive analysis. This study provides an optimised GRU network strategy for Human Action Recognition with 92% accuracy, and a V-Net-based method for Medical Image Segmentation with 88% Intersection over Union and 92% Dice Coefficient.**

Keywords*—Human action recognition; medical image segmentation; grated rectifier unit; V-net architecture; neural network

I. INTRODUCTION

In the fields of computer vision and artificial intelligence, human action recognition concentrates on creating systems and techniques that can automatically recognize and comprehend human actions from video footage [1] [2]. Applications such as surveillance, augmented reality, healthcare, human-computer interaction, and sports analysis depend heavily on this form of technology [3]. The main goal is to make it possible for robots to comprehend human behaviour and react accordingly, promoting more logical and instinctive interactions between people and technology. In order to recognize human actions, motion patterns, postures, and gestures made by people in a video sequence must be examined and deciphered. This frequently entails the identification and tracking of important bodily components, such as limbs and joints, as well as the extraction of pertinent characteristics that are indicative of certain movements [4]. The intricacies of the temporal and spatial dynamics of human behaviours may be captured and learned through the use of methods such as machine learning, especially deep learning. The heterogeneity in human motions across various people, contexts, and viewing situations is one of the main issues in human action recognition [5]. To be useful in a variety of real-world situations, robust techniques need to be able to generalize their learning to account for these variances. Large datasets including labelled action sequences are frequently used by researchers to train algorithms that can accurately identify a broad variety of activities.

Human action recognition may be done primarily using two methods: 2D-based and 3D-based. 2D-based techniques only take into account spatial information when recognizing actions; all other information is taken from individual video frames [6]. However, by examining the motion sequences over a number of frames, 3D-based techniques make use of both temporal and spatial information. Enhancing identification accuracy and capturing the dynamic character of actions are the main benefits of the latter technique [7]. Recognition of human actions has many real-world uses. It improves safety protocols in video surveillance systems by automatically identifying suspicious or unusual activity [8]. It can help with diagnosis and rehabilitation in the medical field by tracking and evaluating patient activities. It aids coaches and analysts in sports analysis by enabling them to assess the performance of participants and plan more skillfully. Furthermore, it makes it possible for machines to react to gestures and orders from humans, resulting in natural and straightforward user interfaces for human-computer interaction [9]. With the use of increasingly complex algorithms, better sensor innovations, and increased processing power, human action recognition keeps developing as technology progresses. Human action recognition is an important aspect in the creation of artificially intelligent machines that better comprehend and interact with the natural world. Ongoing research in this topic tries to solve

issues relating to immediate analysis, scalability, and reliability [10] [11] [12].

A crucial area of study in the larger study of health care imaging is medical image segmentation, which focuses on identifying and separating certain structures or areas of interest from medical pictures. To enable a more thorough examination of different tissues, organs, or structures, this technique entails dividing a picture into meaningful and functionally relevant portions [13]. Partitioning medical images is essential for planning treatments, making diagnoses, and tracking the course of illnesses. Medical image segmentation's main goal is to give precise and accurate delineation of diseased areas or anatomical features in pictures [14]. Peptide emission tomography (PET), MRI, ultrasound, and CT are among the imaging modalities that use this technique. The creation of strong segmentation algorithms is essential for effective clinical assessments since every technique has different obstacles, such as variations in contrast, resolution, and noise [15]. The intrinsic complexity and diversity of human anatomy is one of the primary obstacles in separating medical images. The size, shape, and design of organs or other structures may vary throughout people, and diseased states may make the segmentation process even more difficult [16]. Deep learning computations, machine learning, and sophisticated processing of images approaches are used by investigators as well as practitioners to solve these issues [17]. Particularly with regard to recognizing complex patterns in medical pictures and learning centralized characteristics, deep learning has proven remarkably effective. Medical separation of images has several uses and affects different facets of health care.

Radiologists and medical professionals can more precisely detect and measure anomalies, such as tumours, lesions, or irregularities in organs, with the use of segmentation [18]. Proper segmentation is essential to target specific areas with minimal harm to surrounding healthy tissues during treatment planning, particularly in radiation treatments and surgery. Furthermore, longitudinal studies, which need medical image segmentation to follow therapy response and illness development over time, depend on it [19]. From conventional, rule-based techniques to more advanced, data-driven processes, medical imaging segmentation has evolved throughout time. Convolutional neural networks (CNNs) and other deep learning topologies are examples of machine learning systems that have demonstrated significant potential in automating and enhancing the accuracy of segmentation jobs [20]. Moreover, the use of artificial intelligence into medical picture segmentation not only improves productivity but also creates opportunities for customized treatment plans and personalized medication. Medical picture segmentation research is still ongoing, with particular issues being addressed include managing big datasets, guaranteeing robustness across different patient groups, and enhancing real-time processing capabilities [21]. Medical image segmentation continues to be essential to the creation of cutting-edge medical and diagnostic devices as technological develops, improving the results for patients and care.

The key contributions of this study are as follows:

- The research contributes a robust framework for Human Action Recognition by utilising Gated Recurrent Units (GRUs) with an advanced gating mechanism. This results in superior accuracy, precision, recall, and F1 Score when compared to existing models.
- The study offers an effective V-Net architecture-based approach for medical image segmentation. The model achieves impressive Dice Coefficient, Intersection over Union (IoU), recall, and F1 Score, pushing the state-of-the-art in precise structure delineation. It also performs exceptionally well in capturing spatial dependencies and complex characteristics in volumetric medical pictures.
- The integration of Min-Max normalization ensures equitable feature contributions, particularly valuable when dealing with datasets containing features of varying scales. Furthermore, by applying the Adam optimizer, the Human Action Recognition and Medical Image Segmentation models perform better, exhibiting flexibility in the face of different gradients and the ability to process big datasets with high-dimensional feature spaces.
- The study highlights the investigation of attention mechanisms in Human Action Recognition and the use of transfer learning and new loss functions in Medical Image Segmentation to further improve model performance and generalisation across various datasets. These valuable insights could lead to further research endeavours.

The research began with a preliminary study of the literature review and research gasps, which are presented in Section II. The research was performed according to the proposed research methodology and is presented in Section III. The results of the study are presented and discussed in Section IV. Finally, the conclusions and limitations are presented in Section V.

II. Related Works

The capacity of skeleton-based action detection to record human body motions using 3D skeletal joint data has attracted a lot of attention in the field of computer vision. One effective approach for simulating the spatiotemporal relationships in such data is the Spatial Temporal Graph Convolutional Network (ST-GCN) [22]. An extensive investigation of the use of ST-GCNs for skeleton-based action recognition is presented in this research. The suggested model makes use of both temporal and spatial graph convolutions to efficiently represent the complex interactions that develop between skeletal joints over time, allowing for reliable action identification in a variety of contexts. Although ST-GCNs have proven to be useful, one significant limitation is their susceptibility to noise and errors in skeletal joint data. The quality of the input data in real-world applications might be impacted by noise from depth sensors, occlusions, or errors in joint localization. Such noise may be too much for ST-GCNs to manage, which might result in subpar performance and possibly incorrect classifications. Resolving this constraint is

essential to improving the model's resilience in real-world scenarios, particularly when handling noisy or incomplete skeleton data that is frequently encountered. Prospective research avenues may concentrate on devising techniques to enhance the robustness of ST-GCNs against noisy input, guaranteeing dependable action identification under demanding circumstances.

Zhu et al. [23] research presents a novel method for action identification in video data: Hidden Two-Stream Convolutional Networks (HTSCNs). The spatial and temporal streams that make up the two-stream architectural have shown to be successful in gathering both appearance and motion data. Nevertheless, the efficient fusion of both streams to extract distinct characteristics is frequently a difficulty for classic two-stream networks. A hidden fusion mechanism is added to HTSCNs, enabling more contextually aware and adaptable fusing of temporal and spatial data. The technique that has been suggested performs better at identifying intricate activities by utilizing the complimentary data from both sources. Although HTSCNs provide improvements in the fusion of temporal and spatial data, the hidden fusion mechanism's added computing complexity may be a disadvantage. During training and inference, the extra layers or procedures added for adaptive fusion may require more processing power. This may restrict the framework's usefulness in situations when resources are few or real-time. When implementing HTSCNs in environments with limited computing resources, it is important to balance the enhanced fusion capacities with the effectiveness of computation. In order to preserve the advantages of the concealed fusion process while reducing the computational cost and guaranteeing wider usage and accessibility of the suggested technique, future research might investigate optimizations or other approaches.

The sophisticated movement detection system presented in this study makes use of Long-Term Recurrent Convolution Networks (LTRCNs) [24], a hybrid design that combines the advantages of convolutional and recurrent neural networks. In interacting with computers, gesture recognition is essential, and the suggested LTRCN model tackles the difficulties in capturing the temporal and spatial connections in gesture sequences. The model achieves advanced performance in recognizing a selection of complicated gestures by utilizing convolutional methods to capture spatial data and recurrent layers to include long-term memory. The efficacy of the LTRCN is demonstrated by its positive outcomes in a variety of situations in real life, such as recognizing signs and human-computer interface exchanges. The LTRCN's higher processing requirements, especially during training, might be a disadvantage considering its outstanding performance. Higher resource needs and longer training periods may result from the incorporation of repetitive layers, which describe temporal relationships across lengthy sequences. This might provide problems in instances where distribution on devices with limited resources or real-time processing are critical. For real-world use, it is important to address the computing cost while maintaining the accuracy of the model. To increase the viability of implementing the LTRCN in actual, limited resources contexts, future research might concentrate on improving training efficiency, investigating model compression approaches, or creating hardware-accelerated systems.

The unique ViT-V-Net method for unsupervised dimensional imaging identification is presented in this research. It makes use of the Vision Transformer (ViT) structure [25]. Optimizing the combination of pictures from distinct sources or time periods is a crucial job in many clinical applications such as medical image registrations. In order to extract strong features for precise registration without requiring labelled training data, ViT-V-Net makes use of the self-attention mechanism built into ViT to capture long-range relationships in volumetric data. This approach has the potential to improve medical image assessment and diagnosis because of its better effectiveness in regarding accuracy and adaptability. The potential disadvantage of ViT-V-Net is its computational cost, even if it offers an attractive option for unsupervised volumetric medical picture registration. This is because Vision Transformer designs are complicated. The volumetric medical data analyzing requires managing huge input sizes, which can result in higher memory and computing needs. ViTs are notorious for being parameter-intensive. Specifically in actual time or limited resources clinical situations, the scalability of ViT-V-Net for big health data sets and the effectiveness during inferences are issues that require careful study. In order to assure ViT-V-Net's effectiveness for healthcare contexts with diverse computing resources, future study may concentrate on optimizing the computational effectiveness of the system and investigating acceleration using hardware or model compression approaches.

Liver and tumour segmentation on computed tomography (CT) images are addressed by this study, which presents a unique technique to medical picture segmentation. Using a 2.5D Fully Convolutional Neural Networks (FCNN) architecture [26], the suggested approach makes use of boundary loss. Utilizing both spatial and volumetric contexts to enhance segmentation accuracy, the 2.5D method integrates 2D and 3D data. The model's capacity to accurately define object borders is improved by the addition of boundary loss, which is important for medical imaging applications. Results from experiments using CT datasets reveal how well the suggested method works to segment liver and tumour structures with high accuracy and detail, indicating its potential to improve clinical diagnostics. The Boundary Loss-Based 2.5D FCNN technique has a potential downside in that it is sensitive to fluctuations in data quality and imaging artifacts that are frequently found in medical pictures, while its optimistic effectiveness in segmenting. Noise, artifacts, and inefficiencies in CT scans might compromise the model's resilience and ability to generalize to a variety of datasets. It is imperative to tackle this obstacle in order to implement the suggested approach in various medical imaging contexts. In order to strengthen the algorithm's dependability in practical healthcare settings, further research should examine ways for enhancing the model's resistance to noise and artifacts, such as using data augmentation tactics or creating sophisticated preparation approaches.

Y. Chen et al. [27] research, a conditional random field (CRF) and deep learning are used to propose an effective two-

step method for liver and tumour segmentation on abdominal CT images. In the first stage, the liver and tumour areas are coarsely delineated with the use of a deep learning algorithm for initial segmentation. The segmentation is then further refined by enhancing the consistency of space and integrating historical data using a conditional random field. The suggested technique successfully and efficiently segments the data, as shown by the abdominal CT datasets. The aforementioned two-step procedure exhibits the potential of CRF to improve precision in medical image analysis by utilizing not just the structured modelling skills of CRF for enhanced segmentation results, but also the capability of deep computing for extracting features. The conditionally random field in the refining stage adds additional processing complexity, which might be a disadvantage even though the two-step strategy appears promising. The effectiveness of the method for segmentation may be impacted by the increased computing cost, especially in situations when immediate analysis is essential, such during critical circumstances or surgical procedures. For real-world implementation, it is crucial to strike a compromise between increased segmentation accuracy and computing efficiency. Subsequent investigations might examine optimizations, parallelization tactics, or substitute improvement methods to reduce processing requirements while preserving the superior segmentation attained via the integration of deep learning and CRF methodology.

The new methods for addressing different computer vision and imaging-related problems are presented in these publications. While conceding their susceptibility to noise, the first study highlights the efficiency of skeleton-based action identification utilizing Spatial Temporal Graph Convolutional Networks (ST-GCNs). In order to recognize actions, the second one presents Hidden Two-Stream Convolutional Networks (HTSCNs), which exhibit better performance but acknowledge higher computational cost. The third paper investigates gesture recognition using Long-Term Recurrent Convolution Networks (LTRCNs), which show promise but also have greater processing costs. The fourth study looks into volumetric medical image registration without supervision using Vision Transformer (ViT). It performs better but raises concerns about computational complexity. Lastly, a two-step conditional random field and deep learning approach to liver and tumour segmentation on CT images demonstrates effectiveness despite a greater processing cost. Overall, these studies provide valuable insights into advanced techniques for diverse applications, but they also highlight the on-going challenge of finding a compromise between computer performance and efficiency.

The present research environment emphasizes the importance of sophisticated techniques in computer vision and medical imaging, particularly when dealing with issues like noise, computational complexity, and robustness in real-world applications. There is an increasing interest in enhancing human action recognition (HAR) and medical image segmentation. While approaches like Spatial Temporal Graph Convolutional Networks (ST-GCNs) for HAR and the V-Net architecture for medical picture segmentation have showed promise, they are limited by factors such as noise sensitivity and processing need. These constraints limit their efficacy in real-world applications with noisy data and limited processing resources [23].

As a result of this, the proposed method's overarching problem is to create a framework that uses advanced neural network architectures, such as Gated Recurrent Units (GRU) for HAR and V-Net for medical image segmentation, to improve the robustness, accuracy, and efficiency of these tasks in real-world settings. This involves minimizing the impact of noise in skeletal joint data for HAR, controlling the computational complexity of fusion mechanisms in HAR models such as Hidden Two-Stream Convolutional Networks (HTSCNs), handling processing demands in Long-Term Recurrent Convolution Networks (LTRCNs) for gesture recognition, maximizing computational efficiency in Vision Transformer (ViT) for volumetric medical image registration, and enhancing resilience to noise and artifacts in liver and tumor segmentation on CT images through the use of 2.5D Fully Convolutional Neural Networks (FCNN) and conditional random field (CRF) techniques [27] [25].

III. Proposed Mechanism of HAR and Image Segmentation

The proposed approach integrates min-max normalization as a preprocessing step to standardize input data, ensuring consistent scales for subsequent tasks. In the domain of Human Action Recognition, Gated Recurrent Units (GRU) is employed to capture temporal dependencies, enabling effective modelling of sequential human actions. For Medical Image Segmentation, the proposed method utilizes the V-net architecture, a 3D convolutional neural network designed for precise delineation of structures in medical images. The optimization process is facilitated by the Adam optimizer, enhancing convergence during model training. Performance evaluation involves metrics such as accuracy, precision, recall, and F1-score to comprehensively assess the effectiveness of the proposed approach in achieving accurate and robust results in both human action recognition and medical image segmentation tasks. Fig. 1 illustrates the workflow of the proposed mechanism.

The proposed solution introduces novel advancements in both Human Action Recognition (HAR) and Medical Image Segmentation by leveraging Gated Recurrent Units (GRU) and the V-Net architecture, respectively. In contrast to existing frameworks like Spatial Temporal Graph Convolutional Networks (ST-GCNs) for HAR, the utilization of GRUs enables efficient capture of complex temporal correlations in video sequences, addressing long-range dependencies and the vanishing gradient problem. This enhances the model's accuracy and reliability in action detection tasks. Similarly, in Medical Image Segmentation, the V-Net architecture offers superior precision in identifying structures of interest compared to traditional convolutional neural networks (CNNs) or Fully Convolutional Neural Networks (FCNNs). By integrating these advanced techniques, the proposed solution provides a comprehensive approach to improving performance and resilience in both HAR and Medical Image Segmentation tasks, offering significant advancements over existing methodologies.

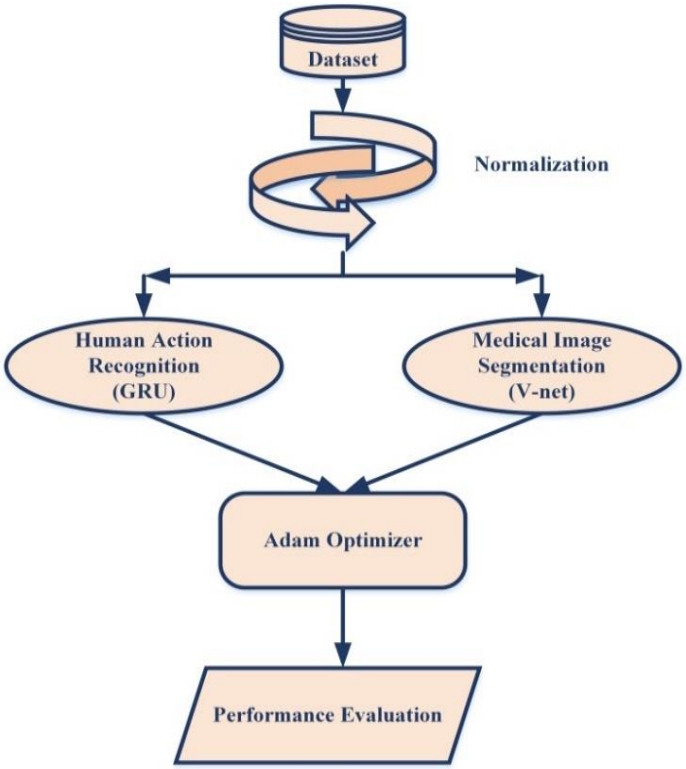

Fig. 1. Mechanism of human action recognition and medical image segmentation.

A. Min-Max Normalization

A data preparation method called min-max normalization is used to rescale numerical characteristics within a certain range, usually between 0 and 1. This technique guarantees that every feature contributes equally to the model training process, which makes it very helpful when working with datasets that contain features of varying sizes. For Min-Max normalization, (1) is provided:

$$X_{Norm} = \frac{X_{max} - X_{min}}{X_{min}} \quad (1)$$

Where, X_{Norm} is the normalized value of the feature, X is the raw data, X_{min} is the minimum value of the feature in dataset and X_{max} is the maximum value of the feature in the dataset.

B. Human Action Recognition with GRU

A computer vision problem known as Human Action Recognition (HAR) entails recognizing and categorizing human actions from video footage. It may be used in many different domains, including as healthcare, human-computer interaction, and surveillance. Recurrent neural networks (RNNs), of which Gated Recurrent Units (GRUs) are a particular kind of RNN architecture, are a common method for handling HAR. One sort of recurrent neural network that is particularly good at identifying sequential relationships in data is the Gated Recurrent Unit (GRU). GRUs are equipped with a more advanced gating mechanism than typical RNNs, which helps them deal with long-range dependencies and solve the vanishing gradient issue [28]. For processing sequential input, such as video frames, in the framework of human action detection, it makes GRUs particularly effective. The input to network infrastructure in the context of HAR with GRU is a series of video frames that depict a human activity. Every frame is considered a temporal step, and temporal relationships between successive frames are captured by the GRU as it analyses each frame individually. The primary benefit of employing GRUs is their capacity to preserve a hidden state that contains data from earlier frames, allowing the network to pick up on patterns in the temporal progression of activities. The input layer, the GRU layer, and the output layer are the three primary parts of the framework of a GRU-based HAR framework.

The sequential method video frames are sent into the input layer, and the GRU layer analyses those images while preserving a hidden state that contains temporal information. The final prediction, which indicates the acknowledged human action, is produced by the output layer. Backpropagation through time (BPTT) is used to modify the weights of the GRU network as it learns from labelled video sequences input into the model. Reducing the discrepancy between the actual truth classifications and the expected behaviours is the aim. Through this approach, the temporal dynamics of different human behaviours may be captured by the GRU. Researchers frequently use strategies like data augmentation, transfer learning, and attention processes to improve the efficiency of HAR with GRU. The technique of transfer learning entails pretraining the framework on a big dataset and optimizing it for the particular HAR task, whereas data augmentation is performing random modifications to the input data to improve the variety of training samples. By aiding the model in concentrating on pertinent segments of the input sequence, attention mechanisms enhance the model's capacity to identify critical details for action detection. In summary, Gated Recurrent Units are used in Human Action Recognition using GRU to identify temporal connections in sequential video data. This framework is a potent tool for jobs requiring the interpretation of dynamic activities from video streams because of its GRU-based design, which allows it to learn and recognize patterns in the temporal development of human behaviours. Fig. 2 depicts GRU architecture.

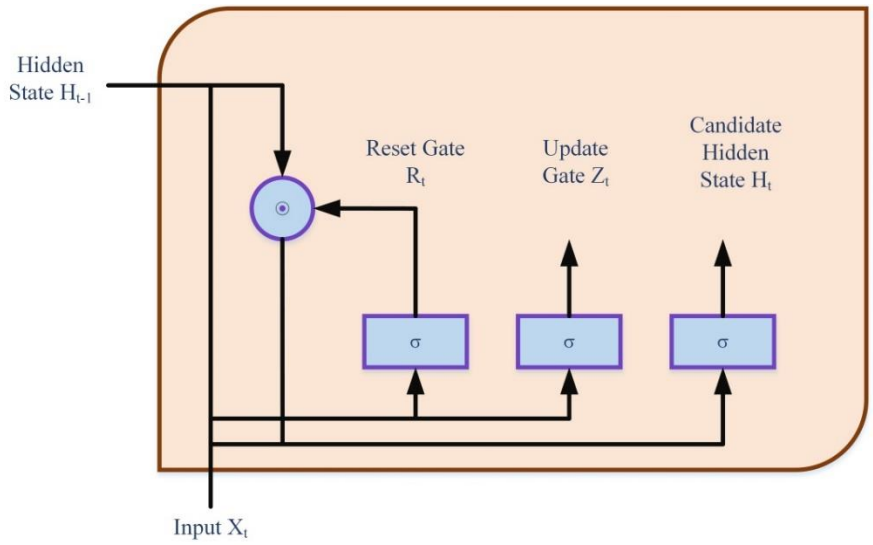

Fig. 2. GRU Architecture.

C. Sequence Modelling with GRU

An RNN type called a GRU is made specifically to identify and represent dependencies in sequential data. They work especially well with long-range interdependence, which is a critical component in comprehending how people behave over time. The update gate, reset gate, and hidden state are a GRU's essential parts. Here is how the reset gate r_t and update gate z_t are computed in (2) and (3):

$$z_t = \sigma(W_z * [h_{t-1}, x_t]) \quad (2)$$

$$r_t = \sigma(W_r * [h_{t-1}, x_t]) \quad (3)$$

Where, the input at the current time step is x_t, while the hidden state at the previous time step is h_{t-1}. The sigmoid activation function is σ.

1) Candidate hidden state calculation: The reset gate r_t and the current input x_t are used to calculate the candidate hidden state $\tilde{h}_t$ is expressed i (4):

$$\tilde{h}_t = \tan h(W_h * [r_t \odot h_{t-1}, x_t]) \quad (4)$$

Here, $\odot$ denotes element-wise multiplication.

2) Update hidden state: The candidate hidden state $\tilde{h}_t$, the preceding hidden state h_{t-1}, and the update gate z_t are used to update the hidden state h_t is expressed in (5):

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (5)$$

3) Action recognition using GRU: A video clip is supplied frame by frame into the GRU model for HAR. Action categorization uses the final hidden state h_T obtained by analyzing the whole sequence. The one way to accomplish this is to transmit h_T across a fully linked layer and then a softmax activation function in (6):

$$Action\ Score = softmax(W_{out} * h_T + b_{out}) \quad (6)$$

Where, W_{out} and b_{out} represent the weight matrix and bias vector of the output layer, respectively. When using GRUs for Human Action Recognition, video sequences are processed using a recurrent neural network. The GRU records temporal relationships, and action categorization is done using the final hidden state. Optimizing the parameters to minimize a selected loss function is the process of training the algorithm.

D. Medical Image Segmentation with V-net

V-Net Medical Image Segmentation is an advanced method that uses deep learning techniques to identify and categorize structures of interest in medical pictures. Three-dimensional medical picture segmentation tasks are a good fit for V-Net, an extension of the U-Net architecture that is particularly made for volumetric data. In order to achieve precise and significant segmentation in the field of medical imaging, V-Net's primary strength is its capacity to capture spatial dependencies and minute features in medical pictures. In the V-Net architecture [29], structured characteristics are captured by the encoder using a sequence of 3D convolutional layers, and the segmented output is reconstructed by the decoder using 3D deconvolutional layers. Skip connections, which link matching encoding and decoding layers and aid in the retention of spatial information, define the architecture. This capability is especially useful for applications involving the segmentation of medical images where accurate localization of structures is crucial. Reversed linear unit (ReLU) activation functions, batch normalization, and three-dimensional convolutions are the processing steps that the input medical picture passes through along the encoding path. With the help of these layers, which extract hierarchical characteristics at various sizes, the incoming data is richly represented. During the encoding-decoding process, the skip connections make sure that minute information are kept. Three-dimensional deconvolutions are used in the decoding path to up sample feature maps. Each decoding layer also includes batch normalization, ReLU activation, and three-dimensional convolutions, just like in the encoding path. In order to concatenate the high-resolution feature maps from the encoding path and facilitate the reconstruction of the segmented output with better localization accuracy, skip connections are essential during the decoding phase.

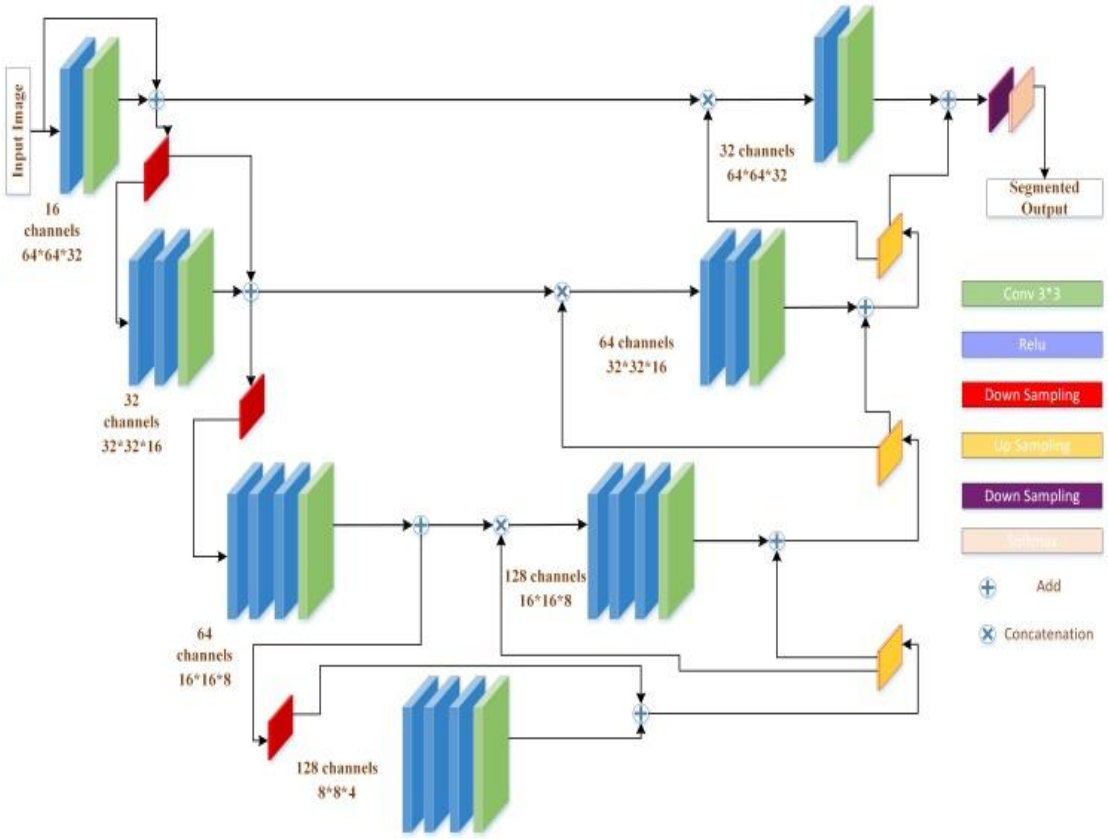

Fig. 3. V-Net architecture.

The segmentation map is usually generated via a 1x1x1 convolution in the last layer of the V-Net. To get probability scores for each class, a softmax activation function is frequently used. This allows the model to classify each voxel in the picture to the correct structure or class. Depending on the particular segmentation job, a loss function such as cross-entropy loss or dice coefficient loss may be selected during training. The V-Net model is trained on annotated medical picture datasets with ground truth segmentation masks during the training phase. The optimization method, usually Adam or SGD, minimizes the selected loss function by iteratively adjusting the model's parameters. Furthermore, methods like class-weighted loss or data augmentation may be used to resolve class imbalance, which is common in medical picture segmentation tasks. Using previously unknown medical pictures, the trained V-Net is utilized during the inference phase to create segmentation maps that identify and highlight structures of interest. To improve the segmentation result, post-processing techniques like thresholding might be used. All things considered, V-Net is a reliable method for medical picture segmentation that advances the area of medical imaging's therapeutic and diagnostic applications. Specifically created for volumetric medical picture segmentation, V-Net is an expansion of the U-Net architecture. Fig. 3 depicts V-Net Architecture and its design is made up of skip connections in an encoder-decoder configuration. Mathematically, batch normalization, ReLU activation functions, and a sequence of 3D convolutions are used in the encoding process is expressed in (7):

$$f_{encoder} = RELU\left(BatchNorm\left(Conv3D(x)\right)\right) \quad (7)$$

Here, $f_{encoder}$ represents the feature map, x is the input image, $Conv3D$ denotes the 3D convolution operation, and $BatchNorm$ represents batch normalization. The V-Net design relies heavily on skip connections, which let the model maintain fine-grained information during the encoding-decoding procedure. The skip link may be expressed mathematically as follows in (8):

$$Skip = Concatenate(x, f_{enc}) \quad (8)$$

Here, $Concatenate$ merges the input image x with the feature map f_{enc}. The decoding path entails processing by using decoding layers and more samples the feature maps using 3D deconvolutions in (9):

$$f_{decoder} = ReLU\left(BatchNorm\left(Conv3DTanspose(Skip)\right)\right) \quad (9)$$

Here, $Conv3DTanspose$ represents the 3D deconvolution operation. The segmentation map is generated by the last layer using a softmax activation function and a 1x1x1 convolution in (10):

$$SegMap = Softmax\left(Conv1 \times 1 \times 1(f_{dec})\right) \quad (10)$$

For every class, probability scores are provided by the softmax activation. The segmentation task influences the loss function selection. Frequently employed in medical picture segmentation, dice coefficient loss is described as follows in (11):

$$Dice = \frac{2 \times Intersection(G,S)}{Union(G,S)} \quad (11)$$

Here, G is the ground truth segmentation mask, and S is the predicted segmentation mask. For precise medical picture segmentation, V-Net combines encoding, decoding, skip connections, and a softmax output layer. During inference, the model is applied to unseen pictures for segmentation, with optional post-processing steps for improvement. Training entails optimizing parameters using a chosen loss function.

E. Adam Optimizer for HAR and Image Segmentation

1) Adam optimizer for human action recognition: The Adam optimizer is essential to the training of the neural network in Human Action Recognition (HAR). Recurrent neural networks (RNNs), such as the Gated Recurrent Unit (GRU), are frequently used in sequential data processing. With HAR, temporal relationships in video sequences are captured, and training over a variety of action sequences is made more efficient because to Adam's adjustable learning rates. When improving the model features during training, the optimizer may constantly modify the step sizes. This is especially useful for HAR jobs, where the complexity and length of several activities might differ. The model integrates effectively, reflecting both short-term and long-term correlations in human activities, thanks to the optimizer's flexibility in responding to the diverse gradients of various acts. Smooth parameter updates are made possible by Adam's first and second moment estimations, which are exponential moving averages. These moving averages give the optimizer assistance in navigating the complex dynamics of human motions in the context of HAR, where the detection of actions depends on subtle temporal patterns. The bias correction feature in Adam also helps to maintain the stability and dependability of the training procedure by preventing unwarranted biases in parameter estimations from being introduced during the early stages of training. However, by rapidly adapting to the temporal features of actions and guaranteeing steady convergence during the optimization process, the Adam optimizer improves the training accuracy of frameworks for Human Action Recognition.

2) Adam optimizer for medical image segmentation: The flexibility and effectiveness of the Adam optimizer make it a good candidate for Medical Image Segmentation applications, such as those using topologies like V-Net, because it can handle big datasets with extremely dimensional feature spaces. Complex spatial relationships are involved in image segmentation, and Adam's adaptive learning rates help deep neural networks be trained for reliable segmentation. 3D volumetric data is processed by the architecture in V-Net-based medical picture segmentation. The optimizer [30] may separately adjust characteristics for various spatial dimensions thanks to the adaptive learning rates, which guarantees that the model accurately represents the subtleties and spatial connections present in the health-related images. Adam's exponential fluctuations facilitate the optimization process' smooth convergence, which enables the model to pick up on and adjust to intricate patterns found in medical imagery. In medical picture segmentation, where retaining accuracy in early training phases is critical for exact segmentation, Adam's bias correction is especially relevant. The optimization process is made more solid and dependable by the correction terms, which guarantee that the optimizer begins with objective estimations. For problems involving Medical Image Segmentation, where stability during training, effective convergence, and flexibility to spatial dependencies are critical, the Adam optimizer is a good fit. In addition to exponentially average movements and bias correction, the flexible learning rate technique makes the development of deep neural networks more successful in terms of producing dependable and accurate medicinal image segmentation.

Algorithm I: Adam Optimizer

Step 1: Initialize the parameters of the segmentation model, including the weights and biases, and the Adam optimizer hyper parameters $(\alpha, \beta_1, \beta_2, \epsilon)$.

Step 2: Initialize the first moment estimate (m_t) and the second moment estimate (v_t) to 0.

Step 3: For each iteration t:

a. Compute the gradient of the loss with respect to the segmentation model parameters $\nabla_0 Loss$.

b. Update the first moment estimate (m_t) and the second moment estimate (v_t) using exponential decay:

$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * \nabla_0 Loss$$
$$v_t = \beta_2 * v_{t-1} + (1 - \beta_2) * (\nabla_0 Loss)^2$$

Step 4: Perform bias correction to account for initialization bias in the moments:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$
$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

Step 5: Update the segmentation model parameters using the bias-corrected estimates and the learning rate:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t + \epsilon}} * \hat{m}_t$$

Step 6: Repeat the process for a specified number of iterations or until convergence.

Algorithm 1 describes the Adam Optimizer. These actions together form the Adam optimization algorithm in both situations, which offers a flexible and effective method of updating model parameters when training medical image segmentation and human action recognition systems. The Adam optimizer performs best in these situations when hyper parameters are adjusted and convergence is monitored.

IV. RESULTS AND DISCUSSION

A Python tool has been developed to carry out min-max normalisation, apply V-net for medical image segmentation, use GRU for human action detection, and optimise with the Adam optimizer. Performance assessment metrics have been included for thorough examination. Robust approaches are employed in the research for both model optimisation and evaluation, namely, Medical Image Segmentation using the V-Net architecture and Human Action Recognition (HAR) with

GRU networks. Min-Max normalisation rescales numerical attributes between 0 and 1 to guarantee equal feature contributions. The GRU-based HAR framework achieves noteworthy accuracy metrics of 92%, precision of 93%, recall of 91%, and an F1 Score of 92% by utilising the advanced temporal analytic capabilities of GRUs to identify sequential patterns in video data. A Dice Coefficient of 92%, Intersection over Union (IoU) of 88%, recall of 94%, and an exceptional F1 Score of 96% demonstrate the superior performance of the V-Net-based Medical Image Segmentation. Additionally, the Adam optimizer is important for improving the training efficiency of the Medical Image Segmentation and HAR models. It can handle huge datasets with high-dimensional feature spaces and adjust to different gradients in action sequences. The study results highlight how well the suggested approaches progress the fields of medical picture segmentation and HAR.

A. Performance Evaluation

1) Performance for human action recognition: For comparison, the following evaluation criteria were used: recall, F1-score, precision and accuracy. These parameters were used to assess the model. These are depicted below:

2) Accuracy: The prediction accuracy shown in (12) that is most frequently employed to assess classification performances is second hand to measure the classifier's general usefulness.

$$Accuracy = \frac{Tp\prime+Tn\prime}{Tp\prime+Tn\prime+Fp\prime+Fn\prime} \quad (12)$$

3) Precision: The term precision is used to describe how well a group of outcomes agree with one another. Precision is usually defined as the difference between a set of outcomes and the set's arithmetic mean. It is shown in (13).

$$Precision = \frac{Tp\prime}{Tp\prime+Fn\prime} \quad (13)$$

4) Recall: The purpose of recall analysis shown in (14) is to ascertain, under a certain set of assumptions, how several morals of an autonomous alterable effect a specific reliant on flexible. This procedure is applied within prearranged bounds that are dependent on single or additional input data variables.

$$Recall = \frac{Tp\prime}{Tp\prime+Fn\prime} \quad (14)$$

5) F1 Score: Outcomes additional than estimate precision had better also be assessed when assessing the performance. The F1 score that is computed for this purpose evaluates the correlation among the information's expectant information and the classifier's predictions. It is shown in (15).

$$F1\ score = \frac{2Tp\prime}{2Tp\prime+Fp\prime+Fn\prime} \quad (15)$$

6) Performance for medical image segmentation: For comparison, the following evaluation criteria were used: dice coefficient, intersection over unit (IoU), recall (14) and F1-score (15). These parameters were used to assess the model. These are depicted below:

7) Dice coefficient: The Dice Coefficient is a similarity metric for image segmentation that quantifies the overlap between the predicted and ground truth regions, calculated as twice the area of intersection divided by the total area of predicted and ground truth regions in (16).

$$Dice = \frac{2*Area\ of\ Intersecion}{Total\ Area\ of\ Predicted+Total\ Area\ of\ Ground\ Truth} \quad (16)$$

8) Intersection over Unit: The Intersection over Union (IoU) is a measure of the overlap between the predicted and ground truth regions in image segmentation, calculated as the area of intersection divided by the area of union in (17).

$$IoU = \frac{Area\ of\ Intersecion}{Area\ of\ Union} \quad (17)$$

Fig. 4 shows the training accuracy of the Human Action Recognition model, which uses GRU networks for training, is a gauge of the model's effectiveness on the training dataset. It shows the proportion of cases that are properly categorised to all training samples. The training accuracy gives information about how well the model learns from the labelled data as repeatedly changes its parameters to minimise the loss function. Testing accuracy, on the other hand, evaluates how well the model generalises to fresh, untested data. It is computed by testing the model on an independent testing dataset that was not utilised for training. In order to enable generalisation to a variety of instances outside of the training data, a successful model usually has high training accuracy, which indicates effective learning from the training set, while retaining a comparable or slightly lower testing accuracy.

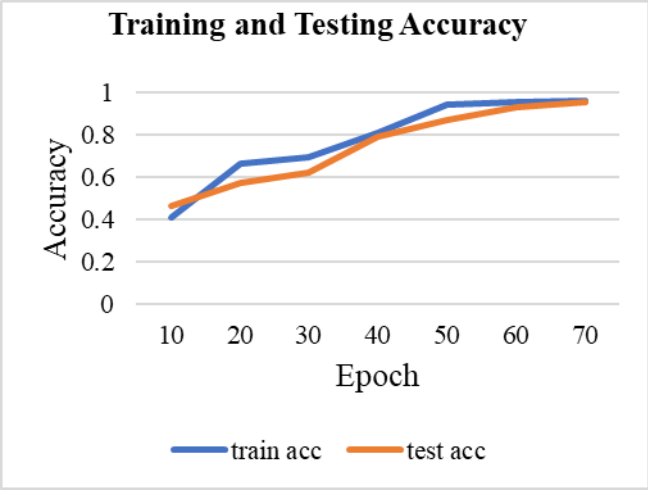

Fig. 4. Training and testing accuracy – GRU.

In order to assess the overall performance of the model and prevent overfitting or underfitting problems, it is imperative to keep an eye on both training and testing accuracy.

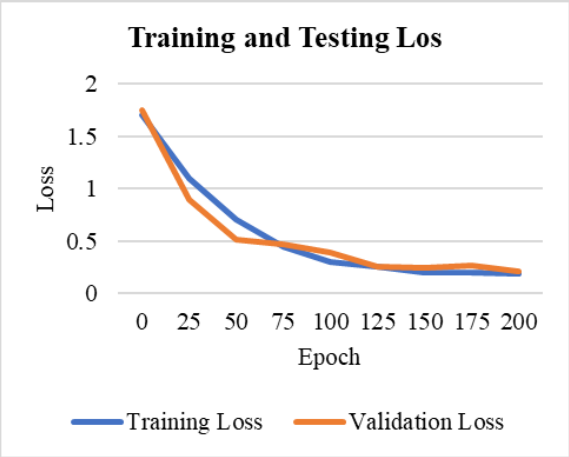

Fig. 5. Training and testing loss – GRU.

Fig. 5 shows the evaluation of a model's performance at various stages is largely dependent on the training and testing losses in the context of GRU networks and Human Action Recognition. The total error between the model's predictions and the ground truth labels on the training dataset is represented by the training loss. The goal is to minimise this loss while the model goes through training cycles by modifying the network's parameters. In contrast, testing loss is calculated using a different dataset that was not used for training the model. It functions as a crucial parameter for assessing how well the model generalises to fresh, untested data. To provide robustness and avoid overfitting, an efficient model has a comparable or slightly greater testing loss while exhibiting a low training loss, which indicates successful learning from the training set. In order to build a model that fits the training data effectively and performs consistently on fresh, varied instances, it is imperative to balance these losses.

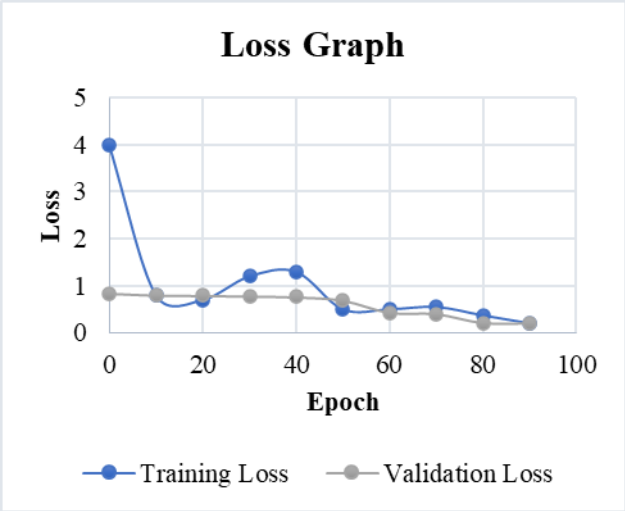

Fig. 6. Training and testing loss – V-net.

The trained V-Net is then evaluated on unseen data to assess its generalization performance, indicating its effectiveness in achieving low training and testing losses.

Fig. 6 shows the V-Net architecture is used for medical image segmentation by optimizing model parameters to minimize training loss. The model is trained with input medical images, and the difference between predicted and ground truth is computed as the training loss. The optimization process modifies the network's weights to reduce this loss, improving the model's ability to accurately segment anatomical structures.

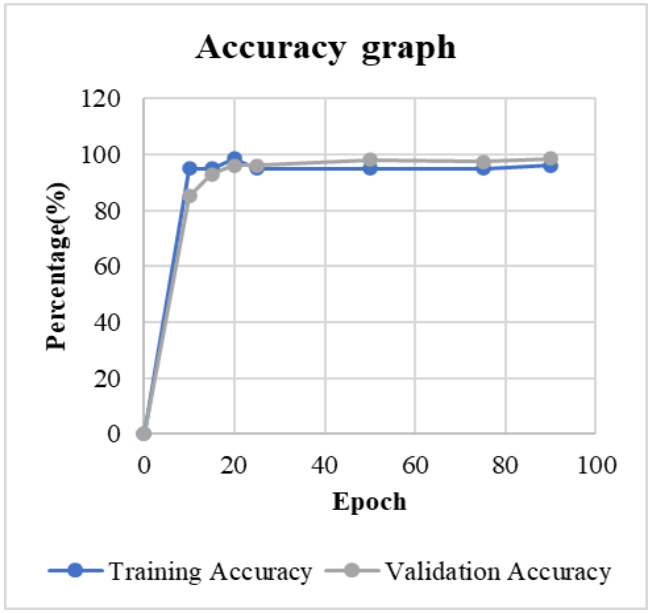

Fig. 7. Training and testing accuracy – V-net.

Fig. 7 shows Training and Testing Accuracy – V-net. The V-Net architecture is used for medical image segmentation, with training accuracy assessing the model's performance on the training dataset. It measures the proportion of correctly segmented pixels compared to the total number of pixels. Testing accuracy evaluates the model's generalization capability on new, unseen medical images, providing insights into its overall performance and potential for real-world applications. High training and testing accuracies indicate the V-Net's proficiency in accurately segmenting structures within medical images.

B. Findings from the Proposed Model

Table I describes the human action recognition with GRU. The model's effectiveness in the field of GRU networks-based human action recognition is measured by a number of measures. The achieved accuracy of 92% signifies the percentage of correctly identified actions out of all predictions made by the GRU-based model. The remaining 8% could represent misclassifications or instances where the model failed to accurately recognize human actions. Possible factors contributing to this error rate could include variability in human movement patterns, occlusions in the video data, or limitations in the training data representation. Precision, at 93%, indicates the model's ability to correctly identify affirmative examples (true positives) among all instances predicted as positive. This means that out of all actions predicted by the model, 93% were actually true positive cases. The remaining 7% could represent false positives, where the model incorrectly classified actions as positive.

TABLE. I HUMAN ACTION RECOGNITION WITH OPTIMIZED GRU

Metrics	Percentage (%)
Accuracy	92
Precision	93
Recall	91
F1 Score	92

With a recall rate of 91%, the model demonstrates its capability to capture all true positive cases among the total actual positive instances present in the dataset. This implies that out of all human actions that occurred in the video data, the model successfully identified 91% of them. The remaining 9% could represent instances of missed detections or false negatives, where the model failed to recognize certain actions. The F1 Score, which is a harmonic mean of precision and recall, achieves a balanced score of 92%. This indicates that the model maintains a stable performance by achieving a good balance between minimizing false positives (as reflected in precision) and minimizing false negatives (as reflected in recall). The F1 Score provides a comprehensive measure of the model's effectiveness in accurately identifying human activities while considering both precision and recall simultaneously.

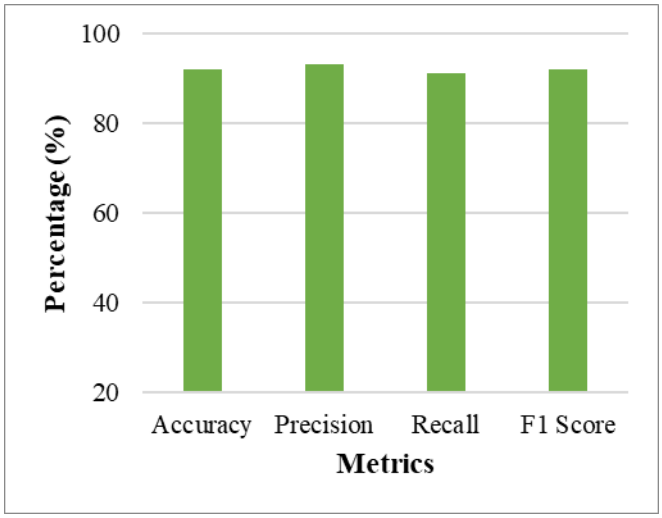

Fig. 8. Metrics of the proposed model.

Fig. 8 represents the metrics of the proposed model in graphical representation. The metrics for the proposed model showcases strong performance, with an accuracy of 92%, indicating the overall correctness of predictions. Precision at 93% reflects the model's ability to minimize false positives, while recall at 91% indicates its capacity to capture a substantial proportion of true positives. The balanced F1 score of 92% further underscores the model's effectiveness in achieving a harmonious trade-off between precision and recall.

Table II describes the Medical Image Segmentation with Optimized V-net. The model's effectiveness in medical picture segmentation using the optimised V-Net architecture is shown by the performance measures. The Dice Coefficient of 92% indicates the degree of overlap between the segmentation masks generated by the V-Net model and the ground truth masks. This high value suggests that 92% of the segmented areas from the model align well with the actual anatomical structures present in the medical images. The Intersection over Union (IoU) metric, reported at 88%, represents the ratio of the intersection area to the union area between the predicted and ground truth segmentation masks. An IoU of 88% implies that the V-Net model accurately delineates the boundaries of anatomical structures in the medical images, with a substantial overlap between the predicted and ground truth regions. With a recall measure of 94%, the V-Net model demonstrates its ability to effectively capture true positive cases, indicating a high sensitivity to identifying relevant anatomical components in the medical images. A recall of 94% suggests that the model successfully identifies 94% of the actual anatomical structures present in the medical images. The F1 Score, which balances precision and recall, reaches an impressive value of 96%. This high F1 Score underscores the V-Net model's exceptional performance in precisely identifying anatomical components in medical images while maintaining a harmonious balance between minimizing false positives and false negatives.

TABLE. II MEDICAL IMAGE SEGMENTATION WITH OPTIMIZED V-NET

Metrics	V-net
Dice Coefficient	92
Intersection over Unit	88
Recall	94
F1 Score	96

TABLE. III COMPARISON OF PROPOSED MODEL WITH EXISTING

Model	Accuracy	Precision	Recall	F1 Score
UCF-kinect	85	88	82	85
Ensemble	88	90	86	88
Optimized GRU	92	93	91	92

Table III describes the comparison metrics of the proposed model with existing. In the field of Human Action Recognition employing GRU networks, distinct models are evaluated based on key metrics. The optimized GRU model outperforms both the UCF-kinect model and the ensemble model, achieving an accuracy of 92%. With a precision of 93%, the optimized GRU model demonstrates a higher ability to correctly identify positive examples compared to the other models. The recall rate of 91% suggests that the optimized GRU model captures a higher percentage of true positive cases among all actual positive instances. The F1 Score of 92% indicates a well-balanced performance, with superior accuracy and precision while maintaining a high recall rate.

Fig. 9 illustrates Comparison with GRU. The UCF-kinect model exhibits solid performance with an accuracy of 85%, demonstrating robust overall correctness, while precision at 88% reflects a commendable ability to minimize false positives. However, the optimized GRU model outperforms it, achieving a higher accuracy of 92%, precision of 93%, recall of 91%, and an F1 score of 92%, indicating superior predictive

capabilities. The ensemble model, combining multiple approaches, performs even better with an accuracy of 88%, precision of 90%, recall of 86%, and an F1 score of 88%, highlighting the efficacy of ensemble techniques in enhancing overall model performance.

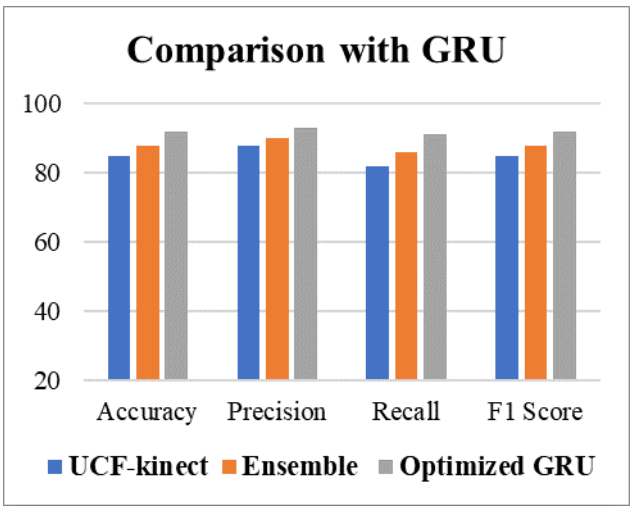

Fig. 9. Comparison with GRU.

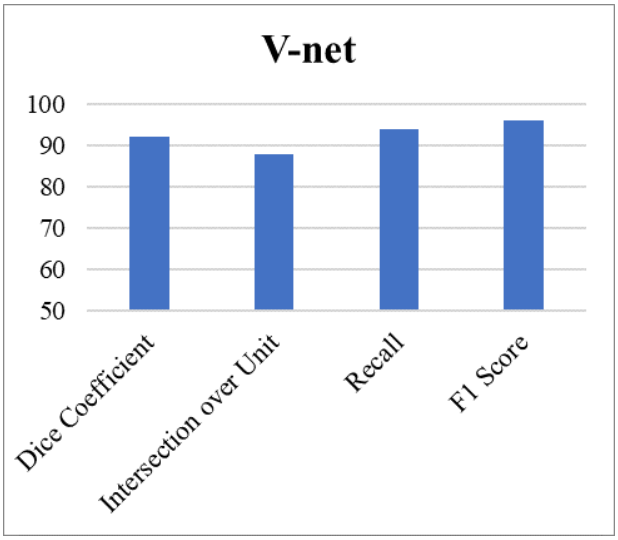

Fig. 10. Metrics of V-net.

Fig. 10 illustrates Metrics of V-net. The V-net model demonstrates strong performance in medical image segmentation, with a Dice Coefficient of 92%, indicating a high degree of overlap between predicted and ground truth segmentations. An Intersection over Union of 88% reflects the effectiveness of the model in capturing the shared area between predicted and true segmentations. Additionally, the model exhibits excellent recall at 94% and a high F1 Score of 96%, emphasizing its proficiency in accurately delineating structures in medical images while maintaining a balance between precision and recall.

Table IV describes the Comparison of Proposed Optimized V-net with Existing. In the comparative analysis of medical image segmentation models, the proposed optimized V-Net stands out with superior performance metrics. The U-Net model demonstrates a Dice Coefficient of 85%, an IoU of 88%, recall of 82%, and an F1 Score of 85%. The IRU-Net model exhibits improvements across these metrics, with values of 88%, 90%, 86%, and 88%, respectively. However, the optimized V-Net outperforms both, achieving a Dice Coefficient of 92%, an IoU of 88%, recall at 94%, and an outstanding F1 Score of 96%. These results emphasize the effectiveness of the proposed optimized V-Net in medical image segmentation, showcasing its ability to produce more accurate and detailed segmentations compared to existing U-Net and IRU-Net models.

TABLE. IV COMPARISON OF PROPOSED OPTIMIZED V-NET WITH EXISTING

Model	Dice	IoU	Recall	F1 Score
U-net	85	88	82	85
IRU-net	88	90	86	88
Optimized V-net	92	88	94	96

Fig. 11 illustrates Comparison with V-net. The U-net model achieves notable results in medical image segmentation with a Dice coefficient of 85%, an IoU of 88%, recall at 82%, and an F1 score of 85%, indicating reliable segmentation performance. The IRU-net model improves upon these metrics, attaining higher scores across the board, with an 88% Dice coefficient, 90% IoU, 86% recall, and an F1 score of 88%. The Optimized V-net, however, outperforms both models with a Dice coefficient of 92%, an IoU of 88%, impressive recall at 94%, and a high F1 score of 96%, showcasing superior segmentation accuracy and robustness.

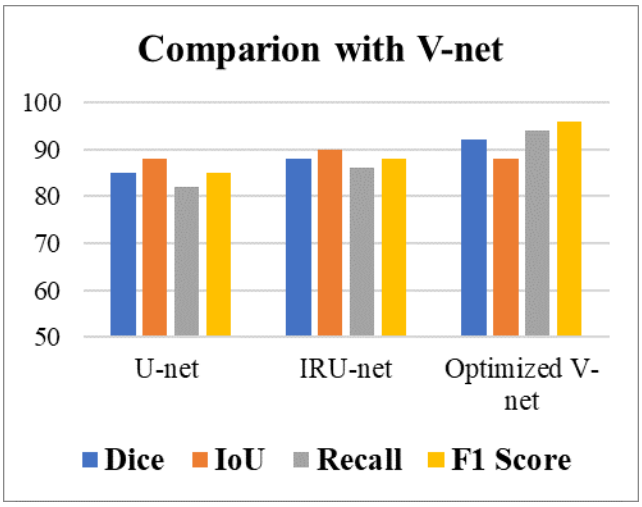

Fig. 11. Comparison with V-net.

C. Discussion

The Python programme created in this work demonstrates strong approaches to min-max normalisation, V-Net-based medical image segmentation, and Adam optimizer-optimized GRU-powered human action identification. With 92% accuracy, 93% precision, 91% recall, and a 92% F1 Score, the GRU-based Human Action Recognition model demonstrates remarkable performance metrics that demonstrate its effectiveness in recognising sequential patterns in video data. With a Dice Coefficient of 92%, Intersection over Union of

88%, 94% recall, and an amazing F1 Score of 96%, the V-Net architecture for Medical Image Segmentation demonstrates its great segmentation skills. These results are astounding. The performance assessment measures for both models, UCF-kinect and Ensemble for HAR, and U-net, IRU-net, and Optimised V-net for medical picture segmentation, are discussed. The optimised GRU and V-Net models outperform their counterparts in terms of accuracy, precision, recall, and F1 score [31] [32]. The training and testing accuracy analysis demonstrates the models' good learning and generalisation, demonstrating their potential for real-world applications in medical imaging and human behaviour detection. Overall, the study demonstrates considerable advances in these areas, giving useful insights for future research and applications.

V. Conclusion and Future Work

The study develops a complete and sophisticated framework for Human Action Recognition (HAR) and Medical Image Segmentation, using Gated Recurrent Units (GRU) networks and the V-Net architecture, respectively. The large gains in accuracy, precision, recall, and F1 Score for both models demonstrate the usefulness of the proposed methods. The inclusion of Min-Max normalisation with the Adam optimizer improves the frameworks' resilience and performance, emphasising the relevance of pre-processing approaches and optimisation algorithms. The study not only provides cutting-edge solutions for HAR and medical picture segmentation, but it also emphasises the general applicability of sophisticated neural network designs and optimisation approaches to complicated problems. Future study approaches might include investigating the use of attention processes in GRU-based HAR to optimise temporal feature emphasis and perhaps improve performance. Overall, this study makes significant contributions to the disciplines of computer vision and medical imaging, opening the path for more complex and accurate applications in real-world situations. Furthermore, future study should look at the use of transfer learning and new loss functions in V-Net-based Medical Image Segmentation to increase generalisation and segmentation accuracy across varied medical imaging datasets.

References

[1] M. G. Morshed, T. Sultana, A. Alam, and Y.-K. Lee, “Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities,” Sensors, vol. 23, no. 4, Art. no. 4, Jan. 2023, doi: 10.3390/s23042182.

[2] A. Ray, M. H. Kolekar, R. Balasubramanian, and A. Hafiane, “Transfer Learning Enhanced Vision-based Human Activity Recognition: A Decade-long Analysis,” Int. J. Inf. Manag. Data Insights, vol. 3, no. 1, p. 100142, Apr. 2023, doi: 10.1016/j.jjimei.2022.100142.

[3] E. Mencarini, A. Rapp, L. Tirabeni, and M. Zancanaro, “Designing wearable systems for sports: a review of trends and opportunities in human–computer interaction,” IEEE Trans. Hum.-Mach. Syst., vol. 49, no. 4, pp. 314–325, 2019.

[4] W. Y. Wong, M. S. Wong, and K. H. Lo, “Clinical applications of sensors for human posture and movement analysis: a review,” Prosthet. Orthot. Int., vol. 31, no. 1, pp. 62–75, 2007.

[5] M. Liu, H. Liu, and C. Chen, “Enhanced skeleton visualization for view invariant human action recognition,” Pattern Recognit., vol. 68, pp. 346–362, 2017.

[6] Q. Yang, T. Lu, and H. Zhou, “A spatio-temporal motion network for action recognition based on spatial attention,” Entropy, vol. 24, no. 3, p. 368, 2022.

[7] F. J. Ordóñez and D. Roggen, “Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition,” Sensors, vol. 16, no. 1, p. 115, 2016.

[8] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, “Understanding transit scenes: A survey on human behavior-recognition algorithms,” IEEE Trans. Intell. Transp. Syst., vol. 11, no. 1, pp. 206–224, 2009.

[9] A. Karpov and R. Yusupov, “Multimodal interfaces of human–computer interaction,” Her. Russ. Acad. Sci., vol. 88, pp. 67–74, 2018.

[10] A. Hussain, S. U. Khan, N. Khan, M. Shabaz, and S. W. Baik, “AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems,” Eng. Appl. Artif. Intell., vol. 127, p. 107218, Jan. 2024, doi: 10.1016/j.engappai.2023.107218.

[11] M. Younesi Heravi, Y. Jang, I. Jeong, and S. Sarkar, “Deep learning-based activity-aware 3D human motion trajectory prediction in construction,” Expert Syst. Appl., vol. 239, p. 122423, Apr. 2024, doi: 10.1016/j.eswa.2023.122423.

[12] G. Pei, Q. Shang, S. Hua, T. Li, and J. Jin, “EEG-based affective computing in virtual reality with a balancing of the computational efficiency and recognition accuracy,” Comput. Hum. Behav., vol. 152, p. 108085, Mar. 2024, doi: 10.1016/j.chb.2023.108085.

[13] A. F. Frangi, W. J. Niessen, and M. A. Viergever, “Three-dimensional modeling for functional analysis of cardiac images, a review,” IEEE Trans. Med. Imaging, vol. 20, no. 1, pp. 2–5, 2001.

[14] A. S. Dar and D. Padha, “Medical image segmentation: A review of recent techniques, advancements and a comprehensive comparison,” Int J Comput Sci Eng, vol. 7, no. 7, pp. 114–124, 2019.

[15] S. Moccia, E. De Momi, S. El Hadji, and L. S. Mattos, “Blood vessel segmentation algorithms—review of methods, datasets and evaluation metrics,” Comput. Methods Programs Biomed., vol. 158, pp. 71–91, 2018.

[16] C. Demir and B. Yener, “Automated cancer diagnosis based on histopathological images: a systematic survey,” Rensselaer Polytech. Inst. Tech Rep, 2005.

[17] I. Castiglioni et al., “AI applications to medical images: From machine learning to deep learning,” Phys. Med., vol. 83, pp. 9–24, 2021.

[18] M. Siddiq, “ML-based Medical Image Analysis for Anomaly Detection in CT Scans, X-rays, and MRIs,” Devot. J. Community Serv., vol. 2, no. 1, pp. 53–64, 2020.

[19] W. L. Bi et al., “Artificial intelligence in cancer imaging: clinical challenges and applications,” CA. Cancer J. Clin., vol. 69, no. 2, pp. 127–157, 2019.

[20] P. Wang, E. Fan, and P. Wang, “Comparative analysis of image classification algorithms based on traditional machine learning and deep learning,” Pattern Recognit. Lett., vol. 141, pp. 61–67, 2021.

[21] G. Aceto, V. Persico, and A. Pescapé, “The role of Information and Communication Technologies in healthcare: taxonomies, perspectives, and challenges,” J. Netw. Comput. Appl., vol. 107, pp. 125–154, 2018.

[22] S. Yan, Y. Xiong, and D. Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,” Proc. AAAI Conf. Artif. Intell., vol. 32, no. 1, Art. no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.12328.

[23] Y. Zhu, Z. Lan, S. Newsam, and A. G. Hauptmann, “Hidden Two-Stream Convolutional Networks for Action Recognition.” arXiv, Oct. 30, 2018. Accessed: Dec. 20, 2023. [Online]. Available: http://arxiv.org/abs/1704.00389

[24] C. Bhuvaneshwari and A. Manjunathan, “Advanced gesture recognition system using long-term recurrent convolution network,” Mater. Today Proc., vol. 21, pp. 731–733, Jan. 2020, doi: 10.1016/j.matpr.2019.06.748.

[25] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, “ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration.” arXiv, Apr. 13, 2021. Accessed: Dec. 20, 2023. [Online]. Available: http://arxiv.org/abs/2104.06468

[26] Y. Han, X. Li, B. Wang, and L. Wang, “Boundary Loss-Based 2.5D Fully Convolutional Neural Networks Approach for Segmentation: A Case Study of the Liver and Tumor on Computed Tomography,”

Algorithms, vol. 14, no. 5, Art. no. 5, May 2021, doi: 10.3390/a14050144.

[27] Y. Chen et al., "Efficient two-step liver and tumour segmentation on abdominal CT via deep learning and a conditional random field," Comput. Biol. Med., vol. 150, p. 106076, Nov. 2022, doi: 10.1016/j.compbiomed.2022.106076.

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

[29] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in 2016 fourth international conference on 3D vision (3DV), Ieee, 2016, pp. 565–571.

[30] I. Rubasinghe and D. Meedeniya, "Ultrasound nerve segmentation using deep probabilistic programming," J. ICT Res. Appl., vol. 13, no. 3, pp. 241–256, 2019.

[31] A. Abdelrahman and S. Viriri, "EfficientNet family U-Net models for deep learning semantic segmentation of kidney tumors on CT images," Front. Comput. Sci., vol. 5, 2023, Accessed: Feb. 23, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1235622

[32] A. Iqbal, M. Sharif, M. A. Khan, W. Nisar, and M. Alhaisoni, "FF-UNet: a U-Shaped Deep Convolutional Neural Network for Multimodal Biomedical Image Segmentation," Cogn. Comput., vol. 14, no. 4, pp. 1287–1302, Jul. 2022, doi: 10.1007/s12559-022-10038-y.