Scientific Research Publishing

Journal of Modern Physics, 2024, 15, 2169-2192
https://www.scirp.org/journal/jmp
ISSN Online: 2153-120X
ISSN Print: 2153-1196

Classifying Vibration Modes Generated by The Michelson Interferometer Using Machine Learning Methods

Xin-Han Tsai[1], Anthony An-Chih Yeh[1], Chen-Hsin Lu[2], Shang-Yu Chou[3], Shih-Wei Wang[4], Chi-Wei Lee[5], Po-Han Lee[1,6*]

[1]The Affiliated Senior High School of National Taiwan Normal University, Taipei City
[2]Department of Materials Science and Engineering, National Tsing Hua University, Hsinchu City
[3]The Department of Dentistry, National Defense Medical Center, Taipei City
[4]Mechanical Science & Engineering Department, University of Illinois at Urbana-Champaign, Urbana, US
[5]Department of Physics, National Tsing Hua University, Hsinchu City
[6]Department of Electro-Optical Engineering, National Taipei University of Technology, Taipei City
Email: *phlee@ntut.edu.tw

How to cite this paper: Tsai, X.-H., Yeh, A.A.-C., Lu, C.-H., Chou, S.-Y., Wang, S.-W., Lee, C.-W. and Lee, P.-H. (2024) Classifying Vibration Modes Generated by The Michelson Interferometer Using Machine Learning Methods. *Journal of Modern Physics*, **15**, 2169-2192.
https://doi.org/10.4236/jmp.2024.1512087

Received: September 25, 2024
Accepted: November 8, 2024
Published: November 11, 2024

Abstract

In this paper, we explore the classification of vibration modes generated by handwriting on an optical desk using deep learning architectures. Three deep learning models—Long Short-Term Memory (LSTM) networks with attention mechanism, Video Vision Transformer (ViViT), and Long-term Recurrent Convolutional Network (LRCN)—were evaluated to determine the most effective method for analyzing time series patterns generated by a Michelson interferometer. The interferometer was used to detect vibration modes created by handwriting, capturing time-series data from the diffraction patterns. Among these models, the LSTM-Attention network achieved the highest validation accuracy, reaching up to 92%, outperforming both ViViT and LRCN. These findings highlight the potential of deep learning in material science for detecting and classifying vibration patterns. The powerful performance of the LSTM-Attention model suggests that it could be applied to similar classification tasks in related fields.

Keywords

Michelson Interferometer, Machine Learning, Vibration Modes, Long Short-Term Memory (LSTM)

1. Introduction

The study of vibration modes generated by materials is essential for applications

in defect detection, material science, and earth science. Various techniques, including optical methods like the Michelson interferometer, have been developed to detect these modes, which produce time-series patterns. However, predicting vibration modes resulting from specific behaviors, such as handwriting, continues to pose a challenging task. The Long Short-Term Memory with Attention mechanism (hereafter referred to as LSTM-Attention) enhances the traditional LSTM model by integrating an attention mechanism, which selectively focuses on and assigns weights to different elements within a sequence. This enhancement allows for better feature extraction and more accurate predictions, particularly in complex and extended sequential data.

LSTM-Attention has proven effective in various applications, such as financial time-series prediction by Xuan Zhang *et al.* [1], speech emotion recognition by Yeonguk Yu and Yoon-Joong Kim [2], and crude oil price forecasting by Hu [3]. Additionally, incorporating the attention mechanism in LSTM models has improved text classification performance [4], and Zhou *et al.* have demonstrated that this model can be applied to cross-lingual sentiment classification [5].

ViViT represents a recent advancement in computer vision, merging the transformer architecture with vision-specific inductive biases. This approach has set new benchmarks in image recognition tasks, positioning it as a promising tool for the classification of vibration patterns. Since the model's release, many researchers have focused on improving its performance, leading to variants such as the Video Swin Transformer and MViTv2 [6] [7]. ViViT's versatility is evident from its applications in areas such as video anomaly detection [8], among others.

The Long-term Recurrent Convolutional Network (LRCN) architecture integrates the feature extraction strengths of Convolutional Neural Networks (CNNs) with the sequential modeling abilities of Long Short-Term Memory (LSTM) networks. This hybrid approach has proven successful in various spatiotemporal tasks, and we explore its potential in classifying vibration modes in this study. LRCN has been applied to a range of challenges. For example, Wei *et al.* demonstrated its effectiveness in early prediction of epileptic seizures [9], and it has been utilized for cyberbullying detection in social media comments [10]. Furthermore, LRCN has been applied to the complex task of handwritten Urdu text recognition, which presents challenges due to its intricacy [11].

The Michelson interferometer is a widely used optical instrument in physics and engineering, known for its precision in detecting minor variations in path lengths by splitting and recombining light beams. This sensitivity allows it to measure various physical phenomena, such as testing general relativity and detecting gravitational waves. In this study, we employ the Michelson interferometer as a detector to generate time series patterns for deep learning-based classification. The Michelson interferometer has proven its utility in applications such as displacement measurements with 10 picometer accuracy [12], detection of third-generation gravitational waves [13], and laser wavelength measurements, as demonstrated by Monchalin *et al.* [14].

Several studies have utilized Michelson interferometers for various purposes, including measuring small displacements, detecting gravitational waves, and material characterization. For instance, Park *et al.* demonstrated its application in vibration measurement [15], while Cheng *et al.* proposed using a Michelson-Sagnac Interferometer to digitize vibration signals, which were then fed into a VGG16 algorithm. Their model achieved an impressive accuracy of 98.44% in a six-class classification task [16].

In addition to the Michelson interferometer, accelerometers are also commonly used to collect vibration data. For example, Medina *et al.* used LSTM to classify signals generated by an accelerometer during gearset operations to detect faults, achieving an accuracy of up to 99.4% across 10 classes [17]. In recent years, machine learning techniques have been increasingly applied to various fields, including physics and engineering. Deep learning methods such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have proven effective in classifying patterns and time series data. For instance, Abdelmaksoud *et al.* used a CNN to classify diverse types of fault signals in induction motors [18], and Yang *et al.* utilized an RNN to predict the remaining useful life of bearings [19].

To the best of our knowledge, no prior research has employed machine learning to classify patterns produced by vibration modes generated by a Michelson interferometer. In this paper, we propose a novel method utilizing neural network architectures to classify vibration patterns produced by handwritten behavior on an optical desk using a Michelson interferometer.

2. Materials and Methods

We propose using three deep learning architectures—LSTM-Attention model, ViViT, and LRCN—to classify vibration modes generated by handwritten behavior on an optical desk. A Michelson interferometer generates a time-series pattern, which is fed into these neural networks, which are all built using TensorFlow and Keras. After training and evaluation, we compare their performances to determine the most effective model. This process is depicted by the Michelson interferometer setup shown in **Figure 1**. This study aims to identify the most accurate model for

Figure 1. The setup for Michelson Interferometer shows the arrangement of components where the interference pattern (green circle) is observed.

this specific task, contributing to fields such as defect detection and material science by providing advanced neural network solutions for classifying vibration patterns.

The computing resources were provided by Taiwan Computing Cloud (TWCC), an AI model training platform developed by the National Center for High-Performance Computing (NCHC). We used TWCC's container computing resources for model training and design, leveraging Nvidia V100 32GB GPUs and adjusting the number of GPU nodes as required.

2.1. Michelson Interferometer Principles

The input beam is split into two by the beam splitter: 50% of the light is reflected toward mirror M_1, while the remaining 50% is transmitted toward mirror M_2. After being reflected from M_1 and M_2, half of the light from each mirror is redirected by the beam splitter toward the viewing screen as shown in Figure 1 and Figure 2.

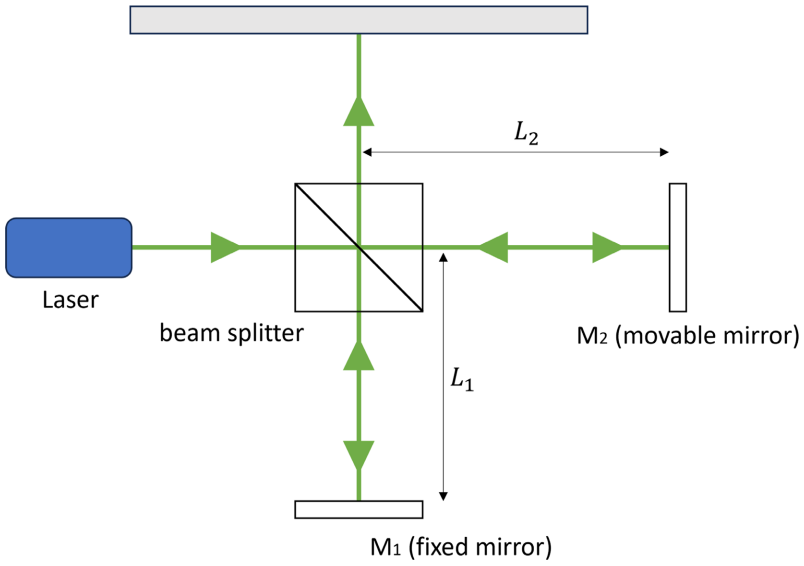

Figure 2. A diagram of a Michelson interferometer showing the light source, beam splitter, mirrors (M_1 and M_2), and the viewing screen. The beam splitter divides the light beam into two paths, which reflect off the mirrors and recombine to form an interference pattern on the viewing screen.

Thus, the original light beam splits, and the resulting beams are recombined. Since the beams originate from the same source, their phases are highly correlated. When a focusing lens is placed between the laser source and the beam-splitter, the light ray spreads out after the focal point, producing an interference pattern of dark and bright rings on the viewing screen. For theoretical analysis, we first define the electromagnetic field of the input laser as E, with its strength denoted as E_0.

After passing through the beam splitter, the incident electromagnetic field E_0 is divided into two components: $E_1 = rE_0$, the reflected part, and $E_2 = tE_0$, the transmitted part, where r and t are the reflection and transmission coefficients,

respectively. The output field E_{out} is then the superposition of these two components as they recombine after reflecting from the mirrors. The resulting interference pattern depends on the phase difference between E_1 and E_2, which leads to constructive or destructive interference on the viewing screen.

Mathematically, this can be expressed as:

$$E_{\text{out}} = rE_1 + tE_2 \tag{1}$$

$$r^2 + t^2 = 1$$

If we introduce the phase difference $\Delta\varphi = \varphi_2 - \varphi_1$ between the two beams due to vibrations or other disturbances, the intensity I_{out} of the output will be affected by the interference of the two beams.

The general equation for the intensity of the output is:

$$I_{\text{out}} = \left|E_{\text{out}}\right|^2 \tag{2}$$

Since the electric field components are now influenced by the phase difference $\Delta\varphi$, the expression for the combined electric field becomes:

$$E_{\text{out}} = E_1 + E_2 \mathrm{e}^{i\Delta\phi} \tag{3}$$

Thus, the intensity becomes:

$$I_{\text{out}} = \left|E_1 + E_2 \mathrm{e}^{i\Delta\phi}\right|^2 \tag{4}$$

Expanding this:

$$I_{\text{out}} = \left|E_1\right|^2 + \left|E_2\right|^2 + 2\left|E_1\right|\left|E_2\right|\cos\left(\Delta\phi\right) \tag{5}$$

Assuming $E_1 = rE_0$ and $E_2 = tE_0$, where r and t are the reflection and transmission coefficients mentioned before, the intensity simplifies to:

$$I_{\text{out}} = \left|E_0\right|^2 \left(r^2 + t^2 + 2rt\cos\left(\Delta\phi\right)\right) \tag{6}$$

This equation shows how the output intensity depends on the phase difference $\Delta\varphi$, which is affected by any vibrations introduced into the system. The constructive or destructive interference leads to variations in the intensity depending on $\Delta\varphi$.

If a small signal $\delta(t)$, this experimental test, is input by handwriting on the optical desk, the final output signal $I_{\text{out\_signal}}$ will be:

$$I_{\text{out\_signal}} = I_0 \cos^2\left(\frac{\Delta\phi_0 + \Delta(t)}{2}\right) \tag{7}$$

where $\Delta(t)$ represents the phase vibration generated by the small signal $\delta(t)$, and $\Delta\phi_0$ is the initial phase difference.

2.2. Data Collection

Vibration mode data is collected using a camera recording at 1080p resolution and 60 frames per second. A Michelson interferometer with a 532 nm laser beam is aligned parallel to the optical desk. The laser beam is expanded to form a larger parallel beam, and a plane mirror is placed perpendicular to the beam. A 50/50

beam splitter cube is positioned between the plane mirror and the light source. The interference pattern is generated by combining the two light beams reflected off the plane mirror as they pass through the beam splitter cube, the complete setup is shown in Table 1.

Table 1. The parts of our Michelson Interferometer setup.

equipment name	spec
Optical Breadboards	aluminum alloy, 700 × 400 × 10 mm, M6-25
solid state green laser	Wavelength: 532nm, Power 10 mW, CW Power Stability: <3% @ 25˚C
laser adjuster	2-axis adjustment
mirror adjuster	ψ1", 2-axis adjustment
beam splitter cube adjuster	2-axis adjustment
Aluminized reflector	O1" × 3 mm Thick, R > 95% for 400 - 700 nm
plano-convex lens	ψ2", EFL = 150 mm
plano-convex lens	ψ1", EFL = 50 mm
Nonpolarized beam splitter	50/50, 1" cube

To generate vibration modes, lowercase English letters (a-z) were handwritten on the optical desk, as shown in Figure 3. Each writing instance was recorded as a video, with each video serving as a data sample for training and testing. Additionally, during intervals between handwritten characters, a pattern with no writing behavior was recorded and labeled as "other" or "_". This occurs during the idle time between each letter. Fifty samples were captured for each letter, all written by the same author.

This results in a total of 50 × 26 = 1300 samples for the letters "a-z" and an additional 1300 samples for the "_" label. To prevent overfitting, we selected 50 "other" samples from the 1300 available for training and evaluation. The data was shuffled and split, with 20% reserved for testing, 16% for validation, and the remaining 64% for training.

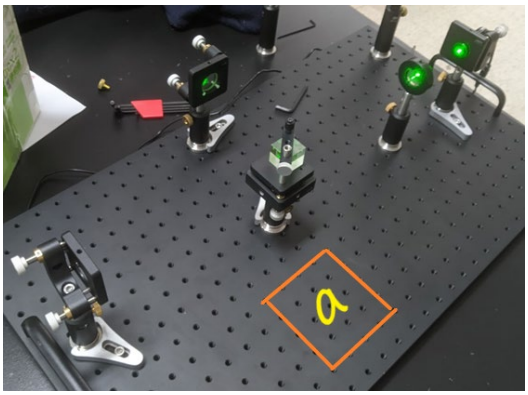

Figure 3. The illustration of behavior of hand-written character "a" on the optical desk.

2.3. Deep Learning Models

In this experiment, the vibration signals are a type of time series data, where the focus is on identifying patterns over time and monitoring continuous steps throughout the entire sequence. Therefore, certain well-known machine learning models that primarily focus on image classification and recognition, such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), as well as deep learning models like Autoencoder + CNN and Autoencoder + Transformer, are not considered in this paper. Instead, we will focus on models such as LSTM (Long Short-Term Memory) with Attention, ViViT (Video Vision Transformer), and LRCN (Long-term Recurrent Convolutional Network), which will be described in detail in the following sections.

The LRCN (Long-term Recurrent Convolutional Networks), as proposed by Donahue *et al.* [20], is a model designed for processing 2D video data by combining the strengths of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). It efficiently processes interferogram frames, with the 2D CNN serving as a feature extractor that captures spatial patterns and relevant information. This network employs convolutional layers with batch normalization to stabilize training, Leaky ReLU activation to enhance feature learning, and max-pooling to downsample the feature maps. Additionally, dropout is applied during training to prevent overfitting.

The spatial features extracted by the CNN are then passed through a flattening layer, converting the 2D representation into a 1D format suitable for the LSTM layer, which manages the temporal patterns in the data. Additionally, batch normalization is applied to assist with optimization, further enhancing the model's structure and performance, as described in [21].

The unique strength of the LRCN model lies in its ability to capture both spatial and temporal dependencies in the interference patterns. This hybrid approach, which combines CNNs for spatial feature extraction and LSTMs for sequential analysis, makes it particularly well-suited for analyzing video data. The model excels at classifying vibration modes by effectively capturing both spatial and temporal patterns, resulting in competitive performance on the interferogram classification task.

The ViViT model, developed by Arnab *et al.* [22], is a state-of-the-art architecture designed for sequence classification tasks using the Transformer's self-attention mechanism. In this study, the ViViT model is applied to 2D video data from interferograms. It leverages multiple attention heads and layers, allowing the model to capture long-range dependencies and interactions within the data. By using self-attention, the ViViT model dynamically focuses on different regions of the interferogram frames, emphasizing the most salient features that are crucial for accurate classification.

The key strength of the ViViT model lies in its ability to capture complex spatial relationships across entire interferogram sequences. By incorporating attention mechanisms, the model outperforms traditional CNN-based methods in capturing

global contextual information, which is essential for accurately classifying vibration modes with intricate spatial patterns. Additionally, the inclusion of label smoothing during training enhances model generalization, resulting in improved accuracy and robustness in the interferogram classification task.

On the other hand, the LRCN model, designed for 1D video data, takes a different approach to handling the concentric circular nature of the diffraction pattern. Instead of processing the entire pattern, the model extracts a single diameter from the interferogram. This simplification preserves essential spatial information while reducing the input data's complexity. The 1D CNN then captures spatial features through layers of 1D convolutions, max-pooling, and dropout, effectively analyzing the one-dimensional time-series patterns of the interferogram.

The main advantage of the LRCN model is its efficiency in handling 1D video data, providing a compact representation while retaining key spatial information. Although it achieves moderate accuracy on the interferogram classification task, the model may face challenges when dealing with more complex spatial patterns due to the simplified input representation.

The LSTM-Attention model [23], on the other hand, is designed to capture temporal dependencies and subtle variations within the 1D time series patterns of the interferogram. Equipped with 128 LSTM units, this model excels at learning long-term dependencies, which is crucial for accurately identifying vibration modes generated by handwritten behavior. The attention mechanism further enhances the model by dynamically focusing on relevant parts of the input sequence, improving its ability to recognize intricate temporal patterns.

Self-attention allows the LSTM-Attention model to dynamically focus on specific parts of the time series patterns, enhancing classification accuracy. This attention mechanism enables the model to adaptively assign weights to different time steps, enhancing its ability to capture and prioritize relevant features critical for accurate classification. Among the models tested, the LSTM-Attention model demonstrates superior performance in classifying vibration modes, achieving higher accuracy. The interpretability analysis of the model further reveals insights into its decision-making process, providing a clearer understanding of how it identifies crucial temporal features for precise classification.

2.4. Utilization of the Models

Each of the three models—LSTM-Attention for 1D, LRCN for 1D/2D, and ViViT for 2D—was trained on preprocessed data using TensorFlow and Keras. During training, categorical crossentropy was used as the loss function, while Adam and RMSProp served as optimization algorithms.

After training, the models were evaluated on a separate testing set. Metrics such as accuracy were calculated to compare the models' effectiveness. A confusion matrix was also generated to visually assess the classification performance of each model.

Comprehensive analysis of the models' results revealed each model's strengths

and weaknesses in classifying the vibration modes generated by handwritten behavior. Additionally, hyperparameter tuning and optimization were applied to improve classification accuracy on the interferogram data.

The findings from this study help identify the most suitable deep learning architecture for classifying vibration modes generated by handwritten movements on an optical desk. The research also contributes valuable insights into applying deep learning for material science and vibration pattern recognition.

2.5. Data Preprocessing

We begin with a 1920 × 1080 RGB video recorded at 60 fps. The video is then converted to grayscale. After that, we crop the video to focus on the concentric circle pattern produced by the interferometer. The cropped video is resized to 175 × 175 pixels and normalize the pixel values by dividing each pixel by 255. This normalization ensures that the data values remain fall the range of 0 to 1, represented as float32. An example of this preprocessing steps is shown in Figure 4.

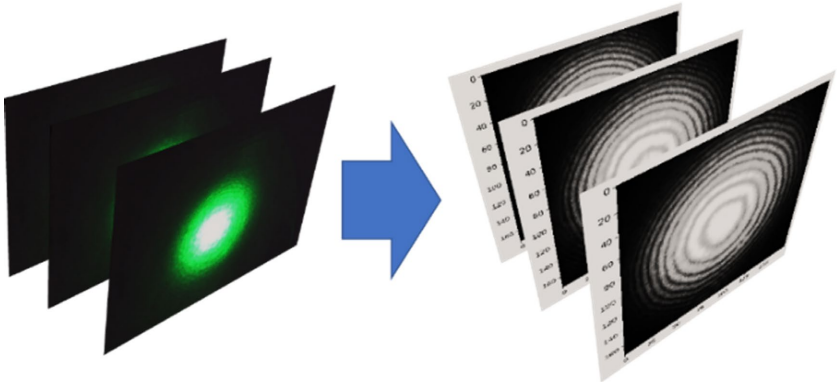

Figure 4. A brief example of the process.

The above-mentioned preprocessing steps were used while training the LRCN 2D model and ViViT. However, since the pattern consists of concentric circles, we realized we could simplify the input data by extracting a single line of pixels from each video frame, passing through the center of the pattern. This extracted line effectively represents the pattern's information. A period of one data sample is shown in Figure 5, where each line corresponds to a single frame of the video.

We can summarize our preprocessing steps as follows, and the workflow is illustrated in Figure 6.

1) Convert 1920 × 1080 @ 60 fps RGB video data to 1920 × 1080 @ 60 fps grayscale.

2) Crop the video to match the concentric circle (700 × 700).

3) Resize it to 175 × 175 using the cv2 default resize algorithm (inter-linear interpolation).

4) Normalize the pixel values to be between 0 and 1 by dividing by 255.

5) Extract the horizontal centerline intensity of the image if using LRCN 1D or LSTM-Attention.

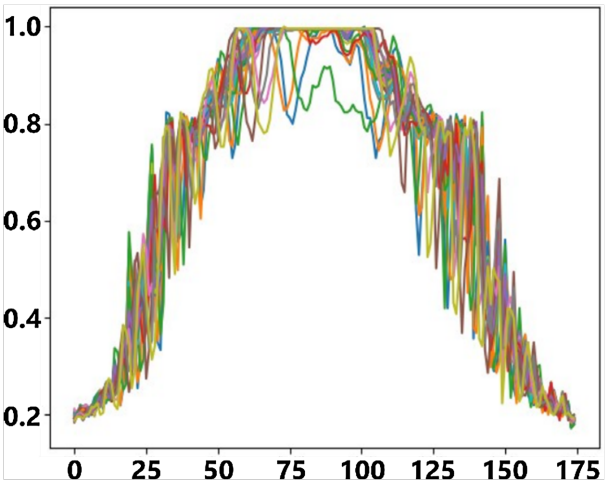

Figure 5. First 40 frames light intensity of an example data.

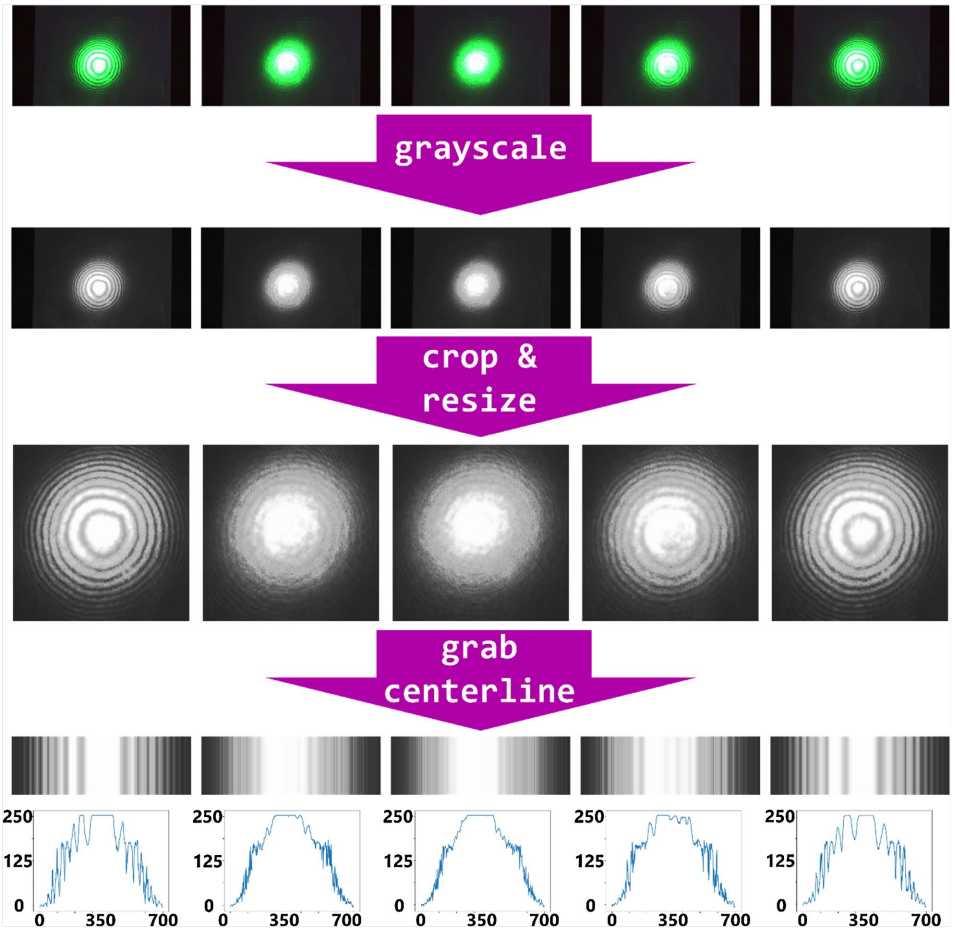

Figure 6. Preprocessing workflow.

The data labeling process utilizes one-hot encoding for 27 classes, which includes 26 English lowercase letters and one additional class for patterns generated during inactivity on the optical desk. Each class contains 50 data points. To prevent overfitting, label smoothing [24] with a factor of 0.1 was applied.

2.6. Model Training

Initially, we trained our model using 2D video data. Since the LRCN model is known for its ability to handle this type of data, we applied it to our dataset. The architecture of the model is illustrated in Figure 7, and the detailed workflow is depicted in Figure 8.

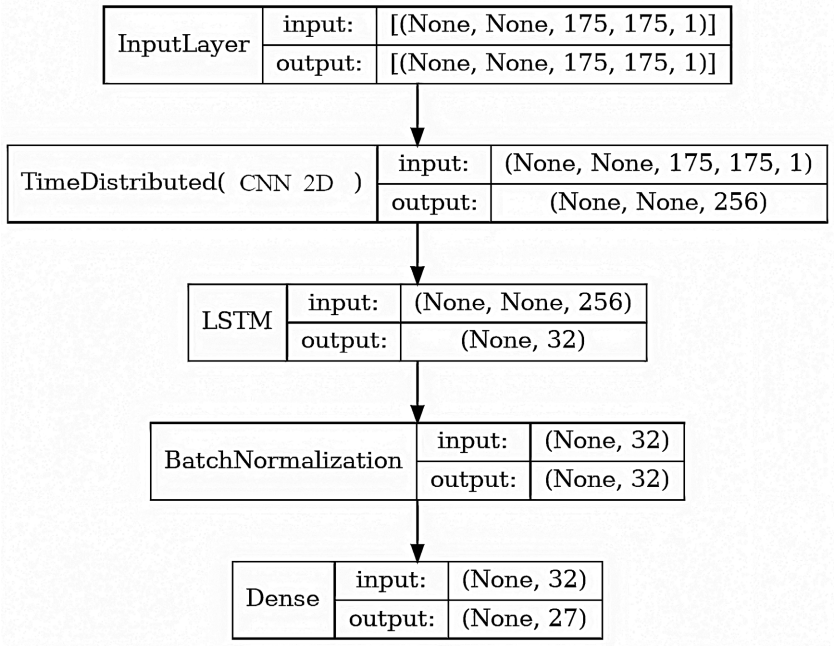

Figure 7. LRCN model architecture.

After applying the LRCN model, we also implemented the ViViT model for comparison. Given ViViT's strength in handling video data through its transformer-based architecture, we explored its performance on our dataset. This allowed us to evaluate the differences in classification accuracy and model efficiency between the two approaches.

There are four variations of the ViViT model proposed by Arnab *et al.* [22]. For simplicity, we focus on evaluating the Spatio-temporal attention variation. The ViViT workflow of our study is shown in Figure 9. The first step involves performing Tubelet Embedding using a conv3D layer. Since conv3D requires a fixed input length, we limit the data length to 175 frames. If the data contains fewer than 175 frames, it is zero-padded, while longer data is cropped.

To improve model performance through data simplification, we applied a feature extraction method that leverages the concentric nature of the diffraction pattern. By extracting only one diameter of the pattern, we preserve the necessary information while reducing complexity.

After completing the feature extraction, we applied the LRCN 1D model, which is simpler than the LRCN 2D model previously used. The architecture of the LRCN 1D model is shown in Figure 10.

The LRCN 1D model shares the same architecture as the LRCN 2D model, with

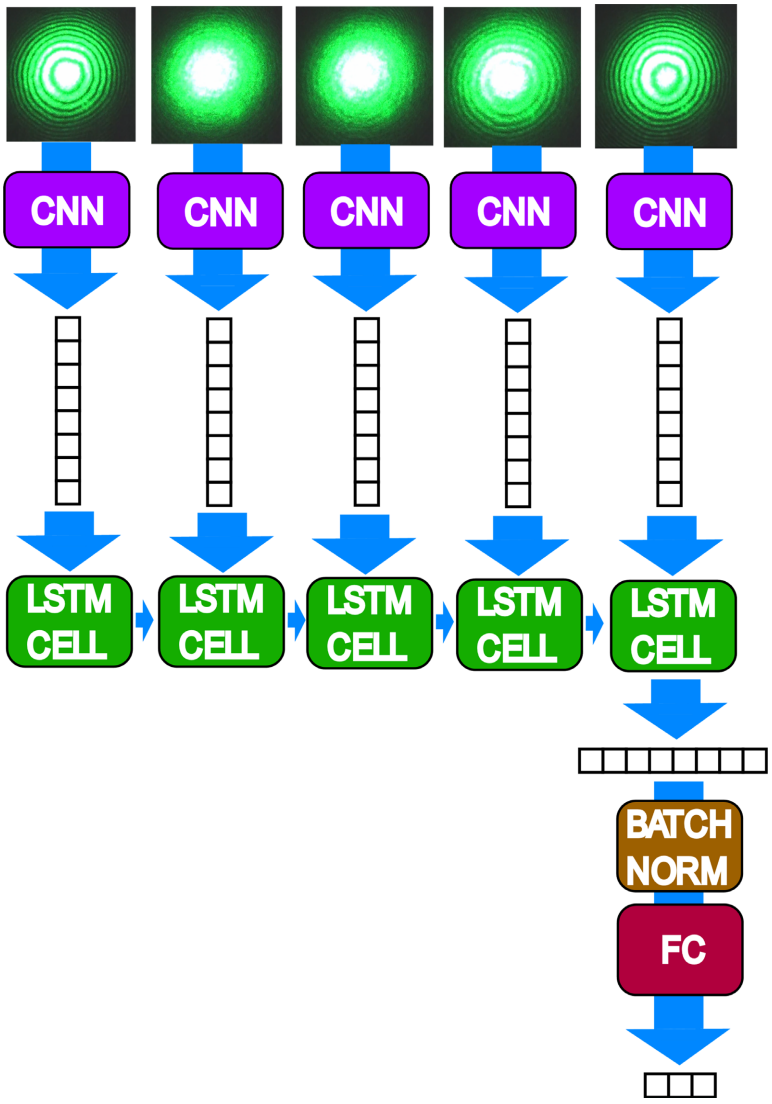

Figure 8. LRCN workflow.

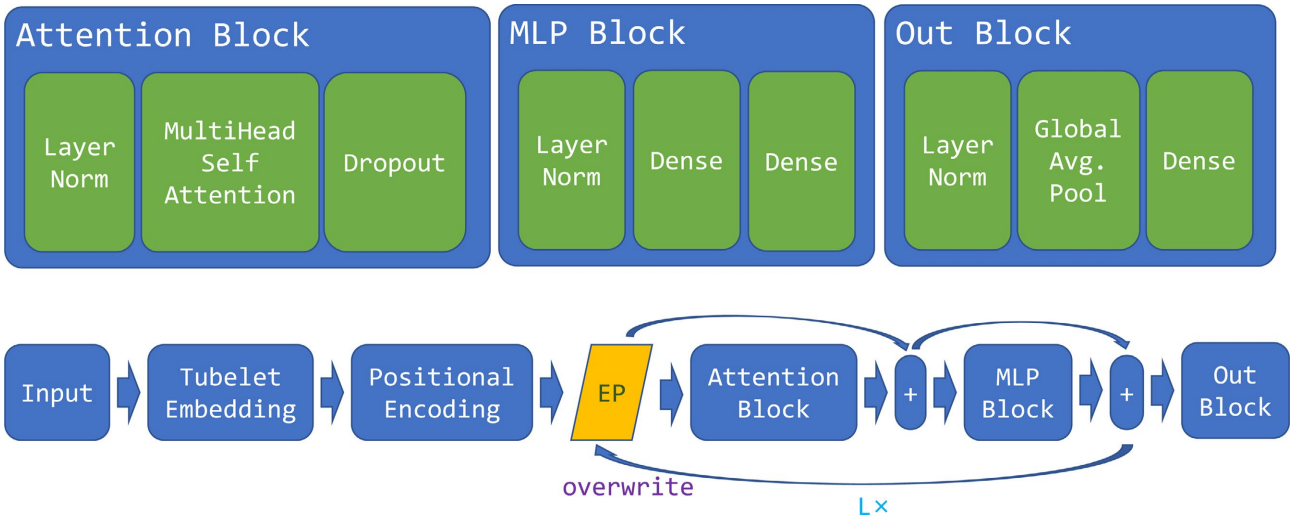

Figure 9. ViViT model workflow.

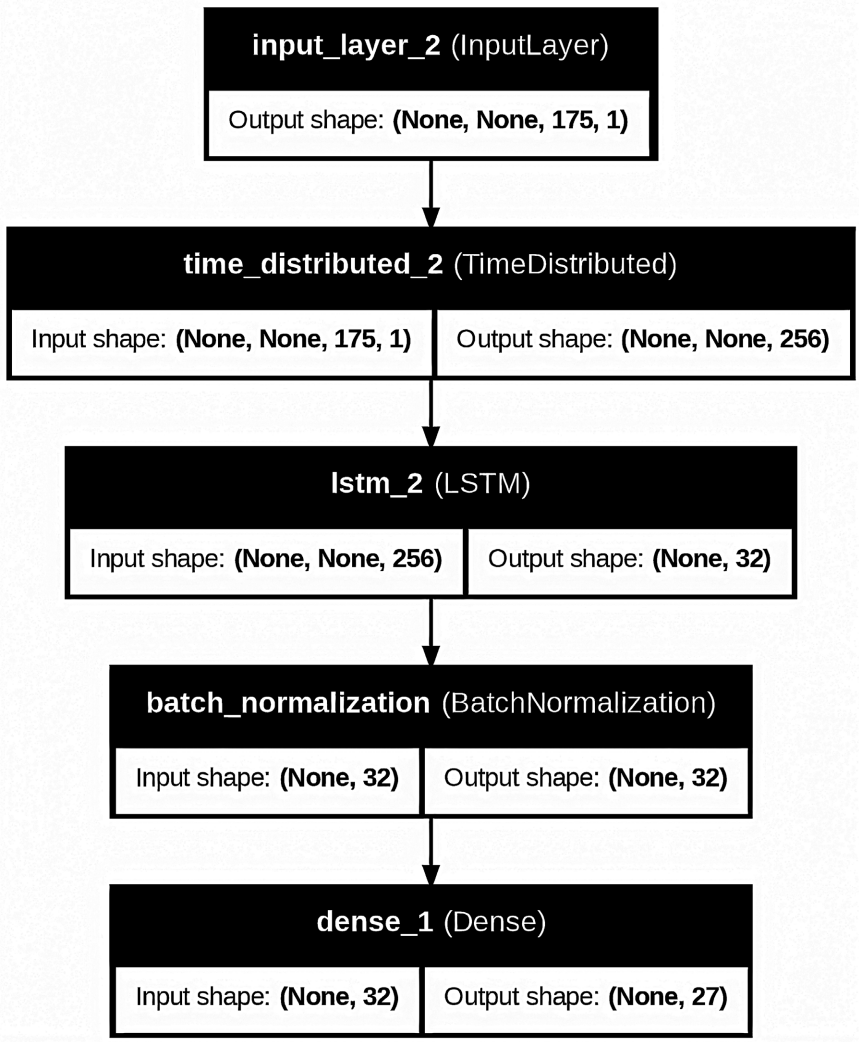

Figure 10. LRCN 1D model architecture.

the key difference being the use of 1D CNN layers for feature extraction instead of 2D CNN layers. In addition, we also explored LSTM-Attention models, which were adapted from the framework proposed by Wang *et al.* [23]. The architecture of the LSTM-Attention model is depicted in **Figure 11**, with the detailed workflow outlined in **Figure 12**. This model utilizes global average pooling, originally proposed by Lin *et al.* [25], which functions similarly to a fully connected layer but with fewer parameters, thereby reducing the likelihood of overfitting. Additional research related to the comparative analysis of speaker identification performance using deep learning can be referenced [26], which provides a strong foundation for the use of multiple classifiers in complex classification tasks.

3. Results and Discussion

Based on our test results, the LRCN model demonstrates optimal performance when configured with four layers of convolutional blocks (Conv -> Pooling -> Batch Norm) and L1 regularization, with the regularization parameter λ set to

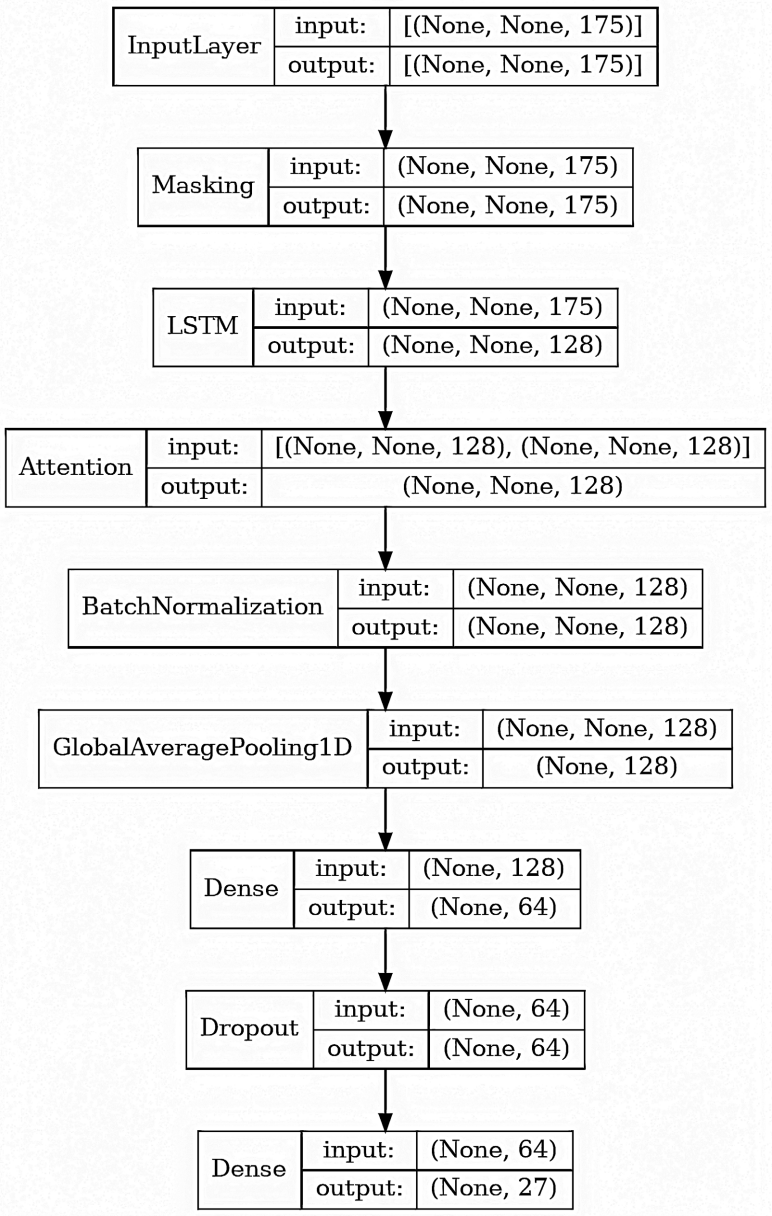

Figure 11. LSTM-Attention model architecture.

0.01. We also used categorical cross-entropy as the loss function. Figure 13 and Figure 14 illustrate the training and validation loss, as well as the accuracy, respectively. These figures provide a clear depiction of the model's performance throughout the training process, reflecting its convergence and ability to generalize to the validation data.

As we can see, the model's validation accuracy is only about 76%, indicating that the model is insufficient for achieving higher precision. The confusion matrix, shown in Figure 15, further highlights the areas where the model struggles, particularly in distinguishing between certain classes.

The confusion matrix reveals that the LRCN model struggles not only with noise but also with distinguishing between similar letters, such as "O" and "N".

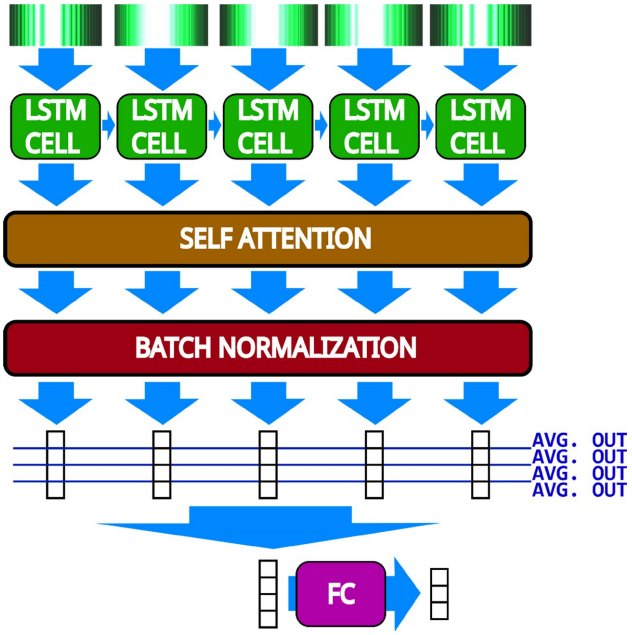

Figure 12. LSTM-Attention workflow.

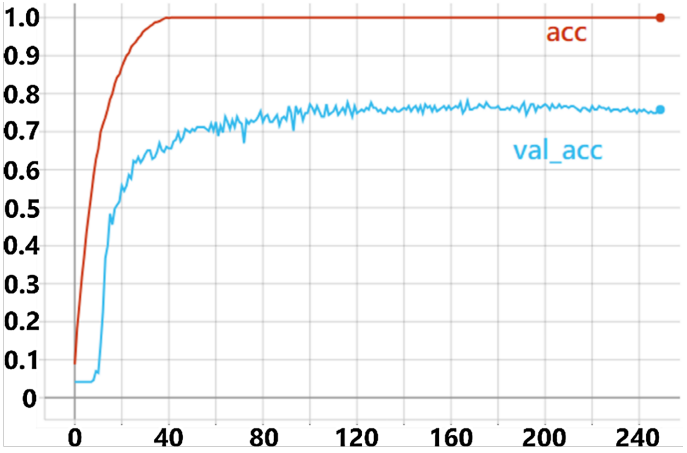

Figure 13. The accuracy graph over epochs by LRCN 2D model training.

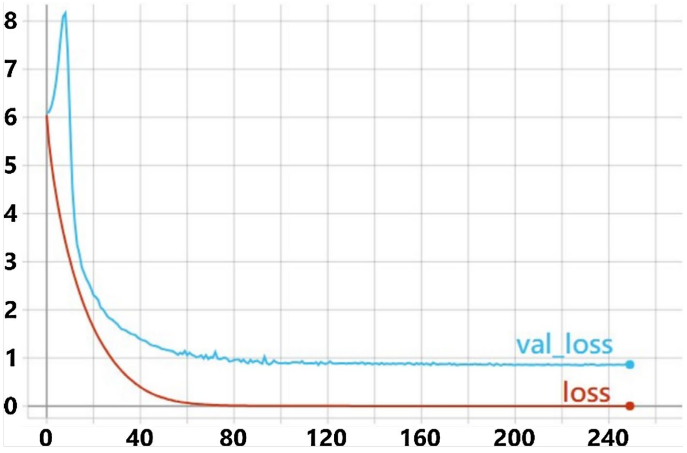

Figure 14. The loss graph over epochs by LRCN 2D model.

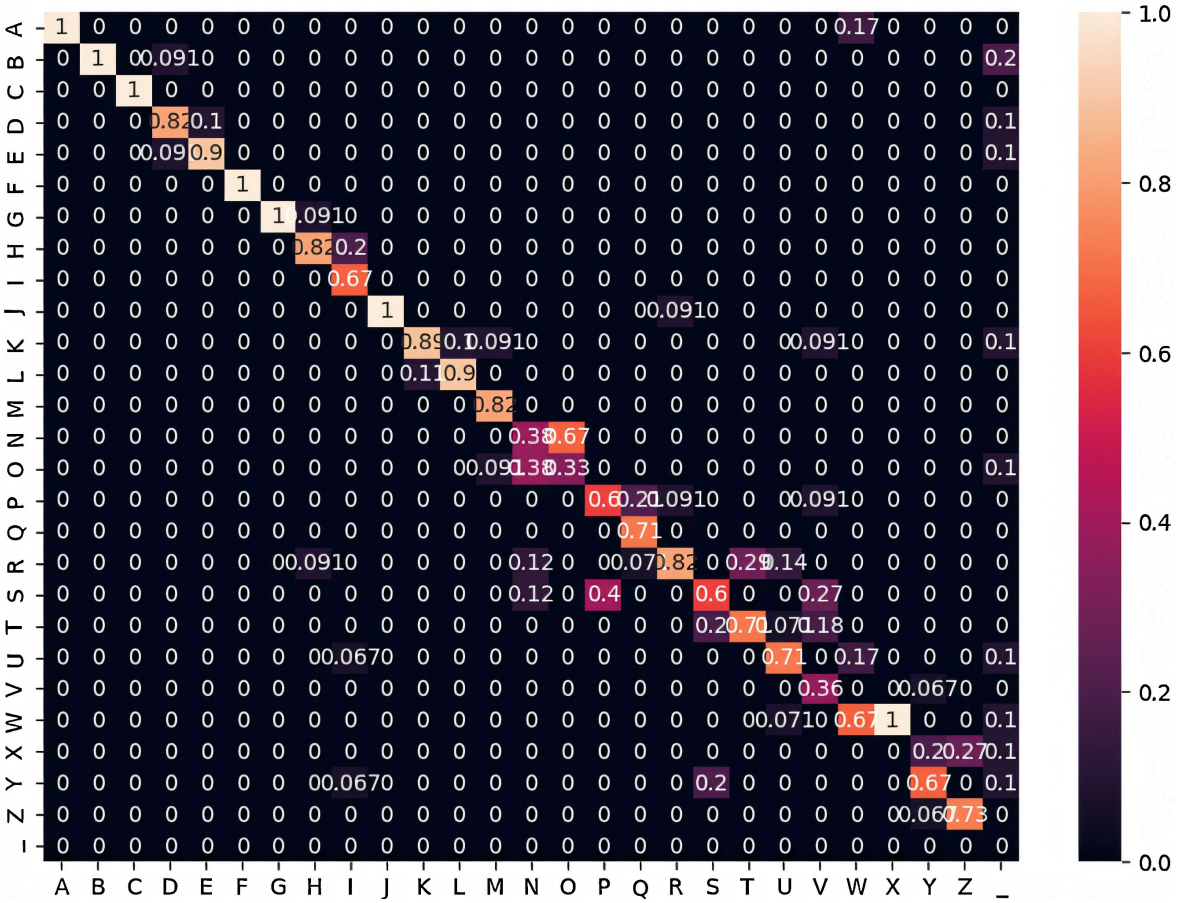

Figure 15. Confusion matrix of LRCN 2D from label a-z and "_".

Additionally, the label "_" remains unidentifiable, possibly due to LSTM's limitations in distinguishing subtle differences in patterns. To improve this, we experimented with the ViViT model, which utilizes the Transformer architecture for sequence classification. Based on our tests, the most optimized configuration for ViViT includes 8 layers of attention (L = 8, as shown in **Figure 9**), with each layer using 8 heads. The patch size is set to (16, 16, 16), and the projection dimension is configured to 27. The training/validation accuracy and loss are depicted in **Figure 16** and **Figure 17**.

As observed, the model achieved a validation accuracy of approximately 77%, but signs of overfitting became evident as the validation loss began increasing around epoch 100. To mitigate this issue, we applied label smoothing, following the recommendation of Arnab *et al.* [22], with a label smoothing factor set to 0.1. The updated training/validation accuracy and validation loss are displayed in **Figure 18** and **Figure 19**.

The best result with ViViT reached a validation accuracy of 86%, which represents a substantial improvement. The loss consistently decreased while the accuracy continued to rise. The confusion matrix, which further illustrates the model's performance, is displayed in **Figure 20**.

The confusion matrix reveals that while the ViViT model outperforms LRCN

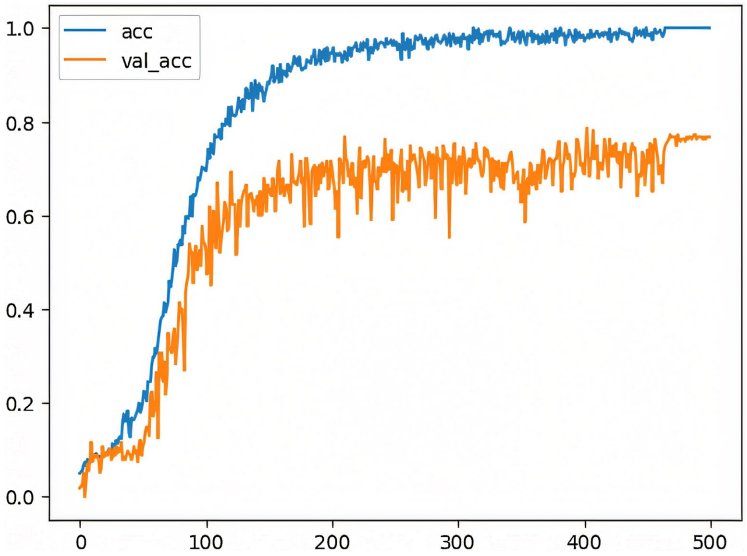

Figure 16. The accuracy graph over epochs by ViViT.

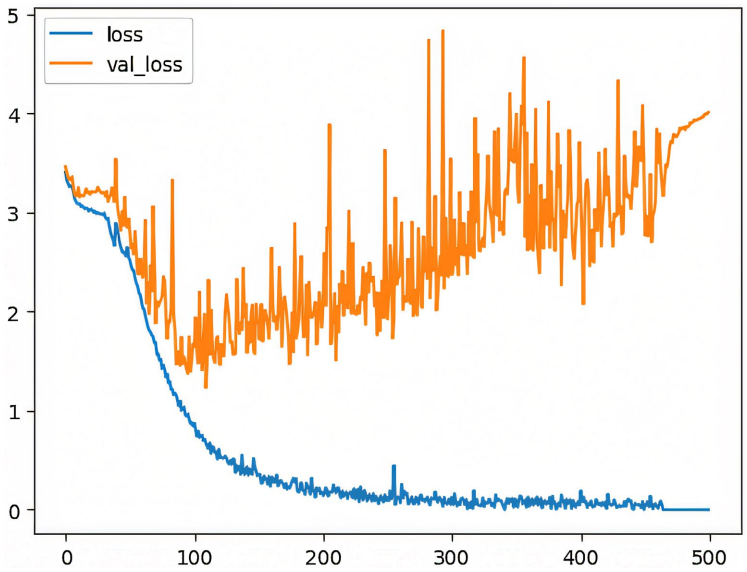

Figure 17. The loss graph over epochs by ViViT.

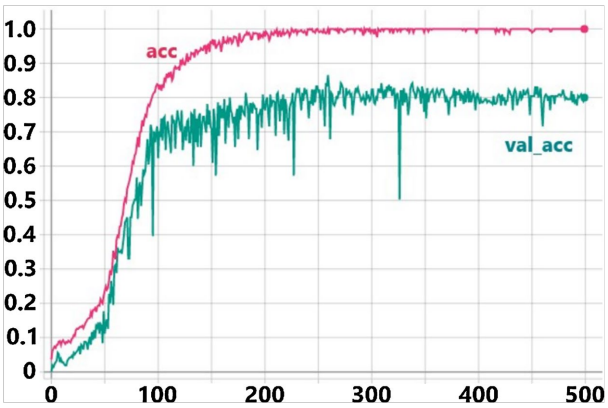

Figure 18. The accuracy graph over epochs by ViViT with label smoothing.

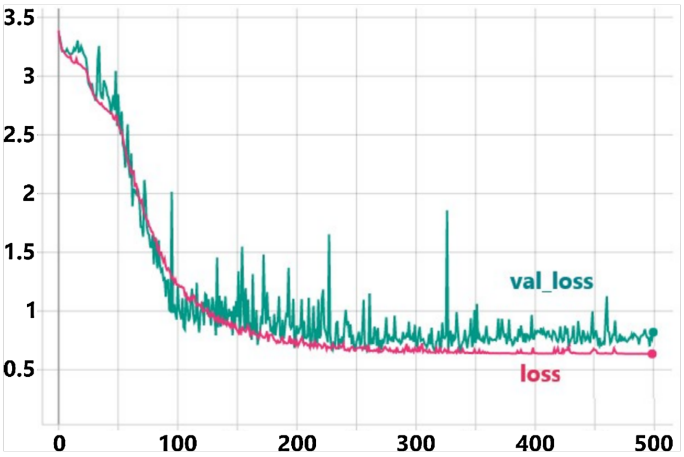

Figure 19. The loss graph over epochs by ViViT with label smoothing.

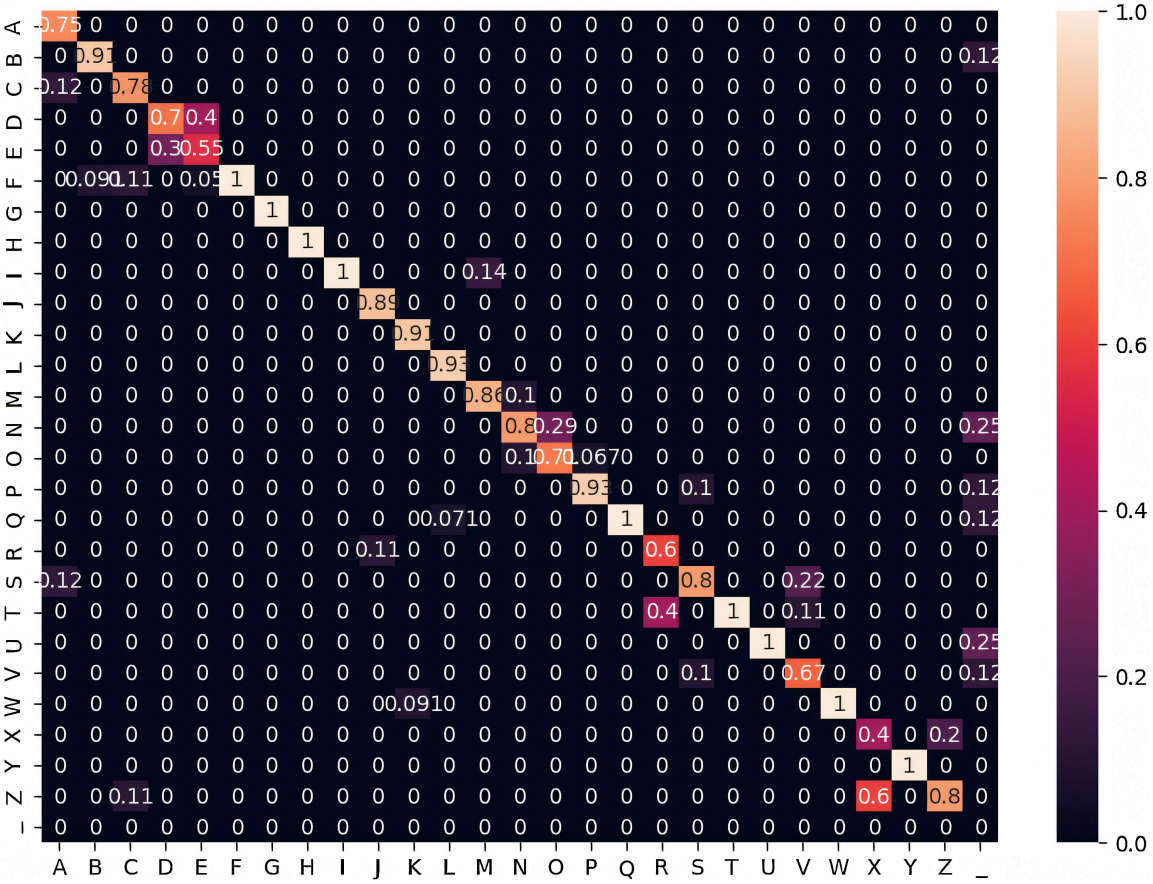

Figure 20. The confusion matrix of ViViT from label a-z and _.

overall, it is not yet delivering optimal performance. The label "_" still shows an accuracy of 0%. We believe there is potential for further improvement. Recognizing that data simplification can often enhance model performance, we explored an innovative approach: since the diffraction pattern consists of concentric circles, we extracted a single diameter from the pattern, which retains the same amount of information. After applying this feature extraction method, we tried the LRCN

1D model due to its simplicity, having previously tested the LRCN 2D. However, the model only achieved 74% validation accuracy, which led us to move forward and try the LSTM-Attention model instead.

The LSTM-Attention model, adapted from Wang *et al.* [23], utilizes global average pooling (first proposed by Lin *et al.*, 2013) to function like a fully connected layer, but with fewer parameters, thereby reducing the risk of overfitting. As anticipated, the model performs well in classifying the dataset, with a progressive decrease in loss value and an increase in accuracy, as demonstrated in Figure 21 and Figure 22.

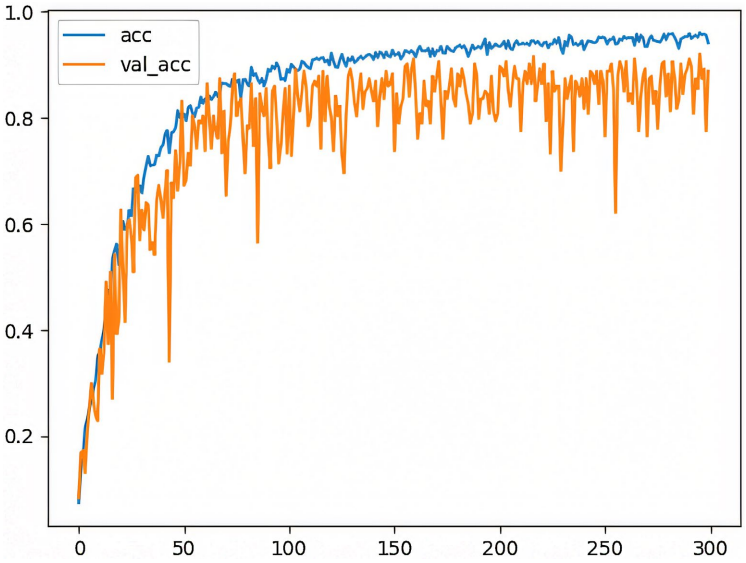

Figure 21. The accuracy graph over epochs by LSTM-Attention model.

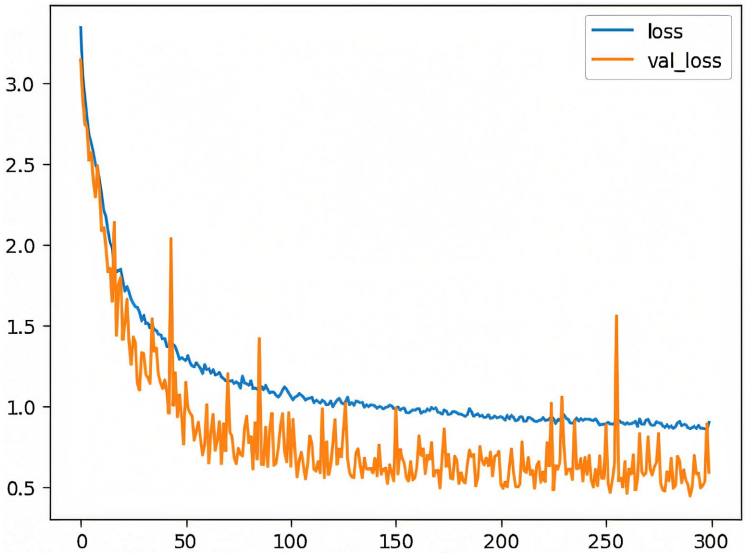

Figure 22. The loss graph over epochs by LSTM-Attention model.

Next, we validated our model using the test dataset, and the resulting confusion matrix is shown in Figure 23. While the model struggles to classify the "_" label

and shows inaccuracy in classifying the "X" label, overall, it performs effectively in classifying most of the data.

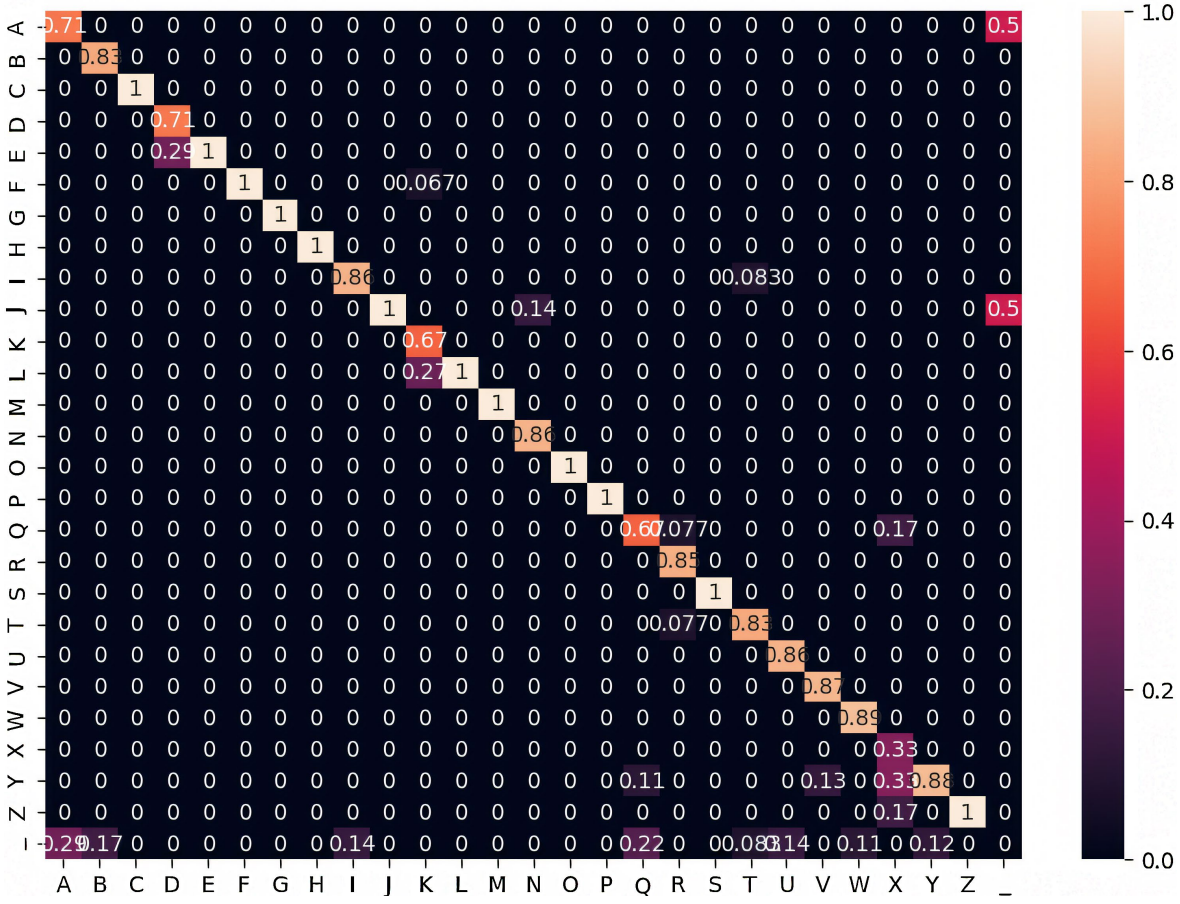

Figure 23. The confusion matrix of LSTM-Attention from label a-z and _.

The best result we achieved was 92% validation accuracy using the LSTM-Attention model. However, the accuracy for the label "_" remains 0%, indicating that the model still struggles with classifying this particular label.

Figures 24-26 illustrate the prediction accuracy for characters a-z and "_" as predicted by the three models: LRCN 2D, ViViT, and LSTM-Attention. Among these, the LSTM-Attention model produced the best results. For the LRCN 2D model, the characters with prediction accuracy below 0.7 include "I, N, O, P, V, S, X". For the ViViT model, the less accurately predicted characters are "E, R, V, X". Meanwhile, for the LSTM-Attention model, the characters "K, Q, X" have lower prediction accuracy. The character "X" proves particularly difficult to recognize across all three models, likely due to the variations in handwriting style and order.

As evident from the confusion matrices of all four models, the label "_"(indicating no writing on the optical desk) proved particularly challenging to classify across all models. This difficulty stems from the fact that the label represents various patterns in each sample, making it hard to categorize consistently. The 0% accuracy for the "_" label may be attributed to the absence of a clear, distinct

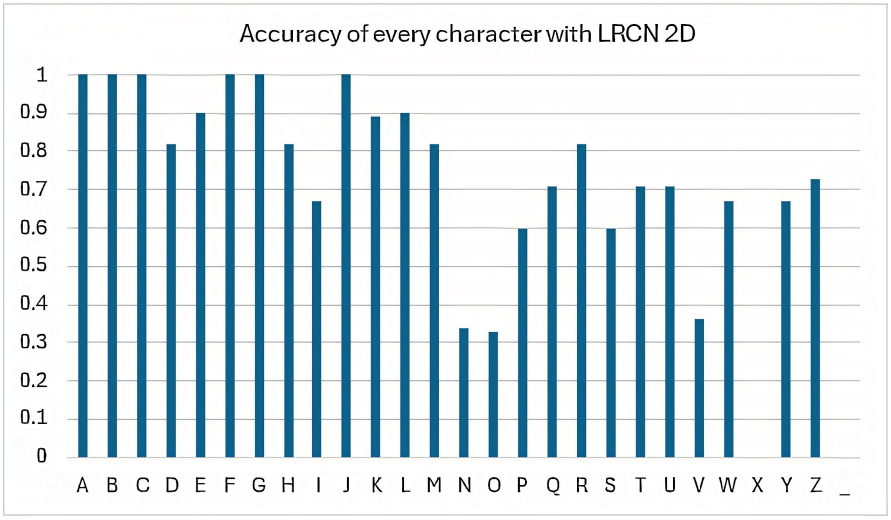

Figure 24. The correctness of character a-z and _ by LRCN 2D model.

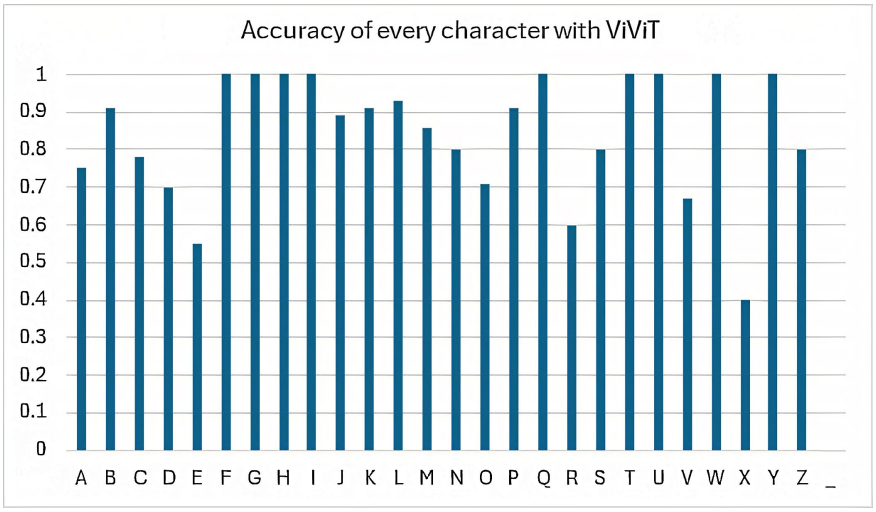

Figure 25. The correctness of character a-z and _ by ViViT model.

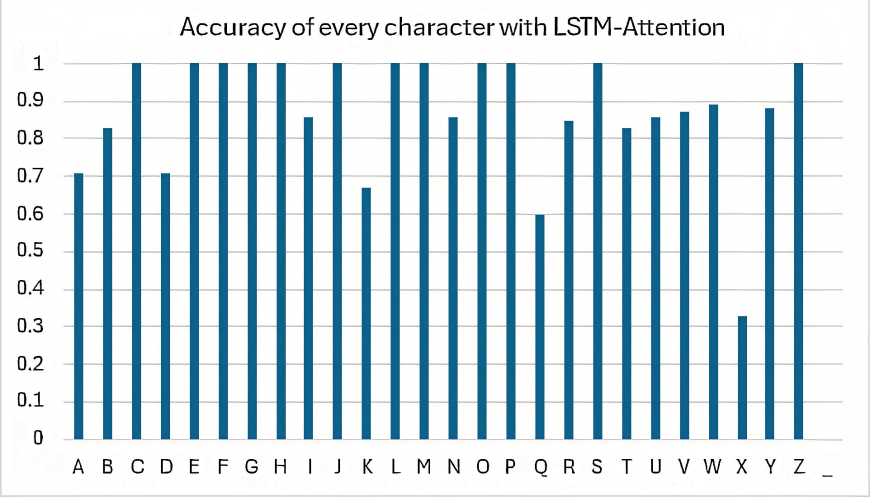

Figure 26. The correctness of character a-z and _ by LSTM-Attention model.

interference pattern, as each sample could contain distinct levels of ambient noise or random signals, resulting in no recognizable vibration pattern for the models.

To address this issue, exploring alternative approaches or further refining data preprocessing techniques is crucial. One potential solution involves augmenting the data to generate more representative patterns for this category. However, when we attempted to train the models with all 1300 samples for the "_" label, we encountered overfitting issues due to the large dataset size, and fine-tuning this extensive data proved too time-consuming to achieve optimal results.

4. Conclusion

Our experimental results demonstrate the successful development of a method that accurately classifies vibration modes using a Michelson interferometer and machine learning techniques. For the vibration signals generated by handwriting, we ranked the performance of three models: LRCN 2D, ViViT, and LSTM-Attention. The LSTM-Attention model excelled, effectively recognizing video data of vibration patterns and achieving an impressive accuracy of 92% in identifying the characters a-z. The system showed strong performance in distinguishing different vibration patterns generated by handwriting on the optical desk, highlighting its potential for precise classification in this context. Beyond handwriting pattern classification, we believe this system has broader potential applications. The ability to recognize various vibration modes suggests that this method could be valuable in fields like earthquake prediction. This research significantly contributes to the understanding of material science and the application of machine learning in vibration pattern recognition. With further research in predictive maintenance, structural health monitoring, and other fields where vibration analysis plays a critical role, these advancements could greatly benefit society.

Acknowledgements

We acknowledge the support provided by the Ministry of Education under Grant MOE A-112-01 (AITC: Promoting AI Education in Elementary and Middle Schools). We also extend our gratitude to the National Center for High-performance Computing (NCHC) for offering computational and storage resources.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Zhang, X., Liang, X., Zhiyuli, A., Zhang, S., Xu, R. and Wu, B. (2019) AT-LSTM: An Attention-Based LSTM Model for Financial Time Series Prediction. *IOP Conference Series: Materials Science and Engineering*, **569**, Article 052037. https://doi.org/10.1088/1757-899x/569/5/052037

[2] Yu, Y. and Kim, Y. (2020) Attention-LSTM-Attention Model for Speech Emotion Recognition and Analysis of IEMOCAP Database. *Electronics*, **9**, Article 713. https://doi.org/10.3390/electronics9050713

[3] Hu, Z. (2021) Crude Oil Price Prediction Using CEEMDAN and LSTM-Attention with News Sentiment Index. Oil & Gas Science and Technology. *Revue d'IFP Energies nouvelles*, **76**, Article No. 28. https://doi.org/10.2516/ogst/2021010

[4] Bai, X. (2018) Text Classification Based on LSTM and Attention. 2018 *Thirteenth International Conference on Digital Information Management*, Berlin, 24-26 September 2018, 29-32.

[5] Zhou, X., Wan, X. and Xiao, J. (2016) Attention-Based LSTM Network for Cross-Lingual Sentiment Classification. *Proceedings of the* 2016 *Conference on Empirical Methods in Natural Language Processing*, Austin, 1-5 November 2016, 247-256. https://doi.org/10.18653/v1/d16-1024

[6] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., *et al.* (2022) Video Swin Transformer. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 3192-3201. https://doi.org/10.1109/cvpr52688.2022.00320

[7] Li, Y., Wu, C., Fan, H., Mangalam, K., Xiong, B., Malik, J., *et al.* (2022) MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 4794-4804. https://doi.org/10.1109/cvpr52688.2022.00476

[8] Yuan, H., Cai, Z., Zhou, H., Wang, Y. and Chen, X. (2021) Transanomaly: Video Anomaly Detection Using Video Vision Transformer. *IEEE Access*, **9**, 123977-123986. https://doi.org/10.1109/access.2021.3109102

[9] Wei, X., Zhou, L., Zhang, Z., Chen, Z. and Zhou, Y. (2019) Early Prediction of Epileptic Seizures Using a Long-Term Recurrent Convolutional Network. *Journal of Neuroscience Methods*, **327**, Article 108395. https://doi.org/10.1016/j.jneumeth.2019.108395

[10] Bu, S. and Cho, S. (2018) A Hybrid Deep Learning System of CNN and LRCN to Detect Cyberbullying from SNS Comments. In: *Lecture Notes in Computer Science*, Springer, 561-572. https://doi.org/10.1007/978-3-319-92639-1_47

[11] Ganai, A.F. and Khursheed, F. (2023) Computationally Efficient Holistic Approach for Handwritten Urdu Recognition Using LRCN Model. *International Journal of Intelligent Systems and Applications in Engineering*, **11**, 536-551. https://ijisae.org/index.php/IJISAE/article/view/2724

[12] Lawall, J. and Kessler, E. (2000) Michelson Interferometry with 10 pm Accuracy. *Review of Scientific Instruments*, **71**, 2669-2676. https://doi.org/10.1063/1.1150715

[13] Freise, A., Chelkowski, S., Hild, S., Pozzo, W.D., Perreca, A. and Vecchio, A. (2009) Triple Michelson Interferometer for a Third-Generation Gravitational Wave Detector. *Classical and Quantum Gravity*, **26**, Article 085012. https://doi.org/10.1088/0264-9381/26/8/085012

[14] Monchalin, J.-P., Kelly, M.J., Thomas, J.E., Kurnit, N.A., Szöke, A., Zernike, F., *et al.* (1981) Accurate Laser Wavelength Measurement with a Precision Two-Beam Scanning Michelson Interferometer. *Applied Optics*, **20**, 736-757. https://doi.org/10.1364/ao.20.000736

[15] Park, S., Lee, J., Kim, Y. and Lee, B.H. (2020) Nanometer-Scale Vibration Measurement Using an Optical Quadrature Interferometer Based on 3×3 Fiber-Optic Coupler. *Sensors*, **20**, Article 2665. https://doi.org/10.3390/s20092665

[16] Cheng, J., Song, Q., Peng, H., Huang, J., Wu, H. and Jia, B. (2022) Optimization of VGG16 Algorithm Pattern Recognition for Signals of Michelson-Sagnac Interference Vibration Sensing System. *Photonics*, **9**, Article 535.

https://doi.org/10.3390/photonics9080535

[17] Medina, R., Macancela, J., Lucero, P., Cabrera, D., Li, C., Cerrada, M., *et al.* (2019) A LSTM Neural Network Approach Using Vibration Signals for Classifying Faults in a Gearbox. 2019 *International Conference on Sensing, Diagnostics, Prognostics, and Control*, Beijing, 15-17 August 2019, 208-214. https://doi.org/10.1109/sdpc.2019.00045

[18] Abdelmaksoud, M., Torki, M., El-Habrouk, M. and Elgeneidy, M. (2023) Convolutional-Neural-Network-Based Multi-Signals Fault Diagnosis of Induction Motor Using Single and Multi-Channels Datasets. *Alexandria Engineering Journal*, **73**, 231-248. https://doi.org/10.1016/j.aej.2023.04.053

[19] Yang, J., Peng, Y., Xie, J. and Wang, P. (2022) Remaining Useful Life Prediction Method for Bearings Based on LSTM with Uncertainty Quantification. *Sensors*, **22**, Article 4549. https://doi.org/10.3390/s22124549

[20] Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., *et al.* (2017) Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 677-691. https://doi.org/10.1109/tpami.2016.2599174

[21] Santurkar, S., Tsipras, D., Ilyas, A. and Madry, A. (2018) How Does Batch Normalization Help Optimization? *Advances in Neural Information Processing Systems*, **31**, 2483-2493.

[22] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M. and Schmid, C. (2021) Vivit: A Video Vision Transformer. 2021 *IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 6816-6826. https://doi.org/10.1109/iccv48922.2021.00676

[23] Wang, Y., Huang, M., Zhu, x. and Zhao, L. (2016) Attention-Based LSTM for Aspect-Level Sentiment Classification. *Proceedings of the* 2016 *Conference on Empirical Methods in Natural Language Processing*, Austin, 1-5 November 2016, 606-615. https://doi.org/10.18653/v1/d16-1058

[24] Müller, R., Kornblith, S. and Hinton, G.E. (2019) When Does Label Smoothing Help? *Advances in Neural Information Processing Systems*, **32**, 4694-4703.

[25] Lin, M., Chen, Q. and Yan, S. (2013) Network in Network. https://doi.org/10.48550/arXiv.1312.4400

[26] Keser, S. and Gezer, E. (2025) Comparative Analysis of Speaker Identification Performance Using Deep Learning, Machine Learning, and Novel Subspace Classifiers with Multiple Feature Extraction Techniques. *Digital Signal Processing*, **156**, Article 104811. https://doi.org/10.1016/j.dsp.2024.104811