FACE MASKDETECTION USING ADEEP LEARNINGAPPROACH

MRS.K.Lavanya *dept. of CSE(AI&ML) Dadi Institute of Engineering and Technology*

E.Hemalatha *dept. of CSE(AI&ML) Dadi Institute of Engineering and Technology*

M.Sowmya *dept. of CSE(AI&ML) Dadi Institute of Engineering and Technology*

B.Pallavi Parimala *dept. of CSE(AI&ML) Dadi Institute of Engineering and Technology*

P.B.S.D.Ganesh *dept. of CSE(AI&ML) Dadi Institute of Engineering and Technology*

Abstract:

The uploaded document contains a project on developing a hybrid Vision Transformer (ViT)-CNN model for face mask detection. Here's a 200word abstract based on the details provided:

This project introduces a novel hybrid Vision Transformer (ViT)-CNN model for efficient face mask detection, leveraging the strengths of convolutional neural networks (CNNs) and pre-trained Vision Transformer backbones. The approach involves local feature extraction using CNN layers combined with the global feature representation provided by the EfficientNetB0 model. The dataset, structured into masked and non-masked categories, is preprocessed with data augmentation techniques to enhance generalization.

The hybrid model integrates CNN-based hierarchical feature maps with transformer-derived global features using concatenation. The architecture includes fully connected layers for binary classification and is trained with an Adam optimizer and binary cross-entropy loss function. Key metrics such as training and validation accuracy are monitored across epochs. The system demonstrates robust performance with a test accuracy of over 90%, validating its applicability in real-world scenarios. Further, the model supports individual image predictions, offering a userfriendly interface for detecting mask compliance. This project emphasizes the synergy between CNN and transformer models in addressing critical healthcare challenges. Future work could explore multi-class classification or real-time deployment for public safety monitoring.

Keywords:

Face Mask Detection, Hybrid ViT -CNN Model, Vision Transformers (ViT),Convolutional Neural Networks (CNNs),EfficientNetB0, Binary Classification .

I. INTRODUCTION

Face masks have become a crucial preventive measure in mitigating the spread of airborne diseases, especially during global pandemics like COVID-19. Ensuring compliance with mask mandates is essential for public health safety, yet manual monitoring is resourceintensive and prone to errors. This has necessitated the development of automated face mask detection systems leveraging advances in artificial intelligence (AI) and computer vision.

Deep learning has emerged as a transformative approach in imagebased classification tasks, enabling high accuracy in recognizing patterns and features. Among deep learning models, convolutional neural networks (CNNs) have proven effective in local feature extraction due to their hierarchical structure and ability to capture spatial information. More recently, Vision Transformers (ViTs) have introduced a paradigm shift by capturing long-range dependencies and global context, traditionally overlooked by CNNs. This project combines the strengths of these two architectures to create a hybrid ViT-CNN model for face mask detection.

The hybrid model leverages CNN layers to extract localized features such as edges and textures while utilizing a pre-trained EfficientNetB0 backbone to capture global feature representations. This dual approach ensures that the model comprehensively analyzes input images, improving classification accuracy. The architecture incorporates fully connected layers for binary classification, distinguishing between images with and without masks.

Data preparation is another critical aspect of the project. The dataset comprises labeled images categorized as “with mask” and “without mask,” and preprocessing steps like resizing, normalization, and data

augmentation are applied to enhance the model's robustness. Training and validation datasets are created using an 80-20 split, ensuring a balanced evaluation.

The system's workflow begins with the extraction of features through CNN layers and the ViT backbone, followed by a concatenation of hybrid features. Fully connected layers subsequently process these features to make predictions. The model is trained using the Adam optimizer and binary cross-entropy loss function over multiple epochs, with accuracy and loss monitored during training.

One of the key objectives is to achieve high performance in terms of accuracy, precision, and recall, making the system viable for real-world applications. A significant emphasis is placed on ensuring the model generalizes well across unseen data, reducing the risk of overfitting. Additionally, the system supports individual image predictions, allowing users to upload images and receive mask compliance results in real-time.

This project addresses a vital public health need by automating mask compliance detection. It combines cutting-edge machine learning techniques to create a scalable and efficient solution applicable in various settings, including healthcare facilities, public transport systems, and workplaces. By merging CNN and ViT architectures, the hybrid model achieves a balance between local and global feature representation, enhancing its classification performance.

Future research could explore expanding the model's capabilities to detect multiple classes, such as incorrect mask usage or other protective gear. Additionally, integration with real-time video feed analysis could further enhance its utility in surveillance and monitoring systems. This project represents a step forward in utilizing AI for societal benefit, showcasing the potential of hybrid deep learning architectures in tackling real-world challenges.

Face masks are vital in preventing the spread of airborne diseases, particularly during pandemics like COVID-19. Monitoring mask compliance in public spaces manually is labor-intensive and prone to errors, highlighting the need for automated solutions. This project presents a hybrid Vision Transformer (ViT)Convolutional Neural Network (CNN) model for efficient face mask detection.

The model combines CNN layers for local feature extraction with a prtrained EfficientNetB0 backbone for capturing global features, offering enhanced accuracy. It utilizes an optimized dataset, processed with augmentation techniques, ensuring robust performance in real-world applications. The system supports binary classification, identifying masked and unmasked faces with high accuracy, and includes real-time image prediction capabilities. This project bridges AI innovation and public health needs effectively

LITERATURE SURVEY

Confront cover discovery has risen as a basic application of computer vision, especially amid the COVID-19 widespread. Various approaches have been investigated in this space, with convolutional neural systems (CNNs) being one of the foremost broadly utilized strategies. CNNs are compelling for extricating localized highlights through convolutional layers and have illustrated victory in early confront cover discovery models utilizing designs such as AlexNet, VGG, and ResNet. Whereas these models exceed expectations in capturing nearby highlights like edges and surfaces, they frequently battle to demonstrate longrange conditions and worldwide designs in pictures. Exchange learning with pre-trained models, counting ResNet50, InceptionV3, and MobileNet, has encourage moved forward execution by leveraging highlights learned from expansive datasets such as ImageNet. This approach decreases the require for broad preparing information and quickens show advancement. In any case, these models still depend intensely on nearby highlight extraction, restricting their adequacy in more complex scenarios. Vision Transformers (ViTs) have as of late picked up consideration for

their capacity to capture worldwide conditions in picture information utilizing selfattention instruments. Whereas ViTs outflank CNNs in scenarios requiring worldwide setting, they are computationally costly and regularly require huge datasets for compelling preparing. To address these challenges, crossover designs combining CNNs and ViTs have developed. These models synergize the qualities of both approaches, empowering viable extraction of both neighborhood and worldwide highlights. Such designs are especially promising for assignments like confront cover discovery, where both point by point designs and by and large setting are imperative. In spite of these headways, challenges stay in accomplishing a adjust between exactness, computational proficiency, and real-time appropriateness. Numerous existing models are not optimized for real-time arrangement, whereas lightweight structures regularly give up exactness. Information expansion and preprocessing methods, such as scaling, turn, and flipping, have been utilized to upgrade demonstrate vigor, but issues like course lopsidedness proceed to influence execution. The proposed cross breed ViT-CNN demonstrate builds on these improvements by combining CNNbased neighborhood highlight extraction with a pre-trained Efficient Net spine for worldwide highlight representation. This approach addresses holes within the existing writing by giving a vigorous and computationally proficient arrangement for confront cover discovery, accomplishing predominant execution in both precision and generalization The proposed framework is built from these advancements since it permits for an all-encompassing arrangement that is through mechanical mediation. more exact, quick, and simple to utilize. It may be a awesome change within the dynamic advancement of physical environment security and security

METHODOLOGY

The Face Mask Detection System is designed to leverage computer vision and deep learning algorithms to create a reliable and efficient model that detects whether individuals in an image are wearing masks. This section elaborates on the methods used in the system, which integrates sophisticated image analysis techniques, deep learning architectures, and real-time prediction capabilities.

*Image Acquisition and Frame Capture *

The system begins by acquiring image data from a curated dataset containing categories such as "with mask" and "without mask." This dataset is extracted from a zip archive, and the images are preprocessed to prepare them for model training. For real-time detection, the system captures frames from live video feeds or uploads, analyzing these frames at regular intervals. Preprocessed frames are standardized in size and scaled to ensure consistency and computational efficiency.

*Image Preprocessing and Feature Extraction *

To features. These steps optimize the input for subsequent classification tasks, ensuring only relevant features are extracted. enhance performance, the images undergo several preprocessing steps. Frames are resized to a fixed size of 128x128 pixels and normalized by scaling pixel values to the range of 0 to 1. Gaussian blur is applied to reduce noise, and grayscale conversion is performed to minimize computational complexity without losing essential

Mask Detection with Hybrid ViT-CNN Model

The core of the system is a hybrid Vision Transformer-Convolutional Neural Network (ViT-CNN) model, designed to capture both local and global features.

- *CNN Layers*: These layers extract localized features such as edges, textures, and contours. Batch normalization and max pooling improve feature extraction and reduce dimensionality.
- *Vision Transformer Backbone*: EfficientNetB0, a pretrained model, captures global dependencies in the images through hierarchical feature extraction.
- *Feature Fusion*: Features from both the CNN and Vision Transformer components are concatenated for a comprehensive representation, enabling robust classification.

The hybrid model is trained using the Adam optimizer and binary crossentropy loss function to ensure efficient and accurate learning. Dropout regularization minimizes overfitting, and the final output layer employs a sigmoid activation function to classify images as "with mask" or "without mask."

Evaluation and Real-Time Prediction

The system is evaluated using validation datasets to monitor performance metrics such as accuracy and loss. Training results are visualized through plots, showcasing trends in model convergence. For real-time deployment, the system integrates OpenCV for image processing and accepts singleimage inputs for instantaneous predictions.

Alert Mechanism and Scalability

Although primarily designed for mask detection, the modular architecture of the system enables scalability and adaptation to other tasks, such as intrusion detection. Lightweight and efficient, the model is deployable on standard hardware, making it suitable for diverse applications, including public health monitoring and security enforcement. The proposed methodology demonstrates a robust and versatile solution for real-world challenges in face mask detection. computational load hence impacts an extensive range of applications by speeding up processing.

Hybrid ViT-CNN Model for Mask Detection

Hybrid ViT-CNN Framework for Veil Location CNN Layers

- Extract localized features such as edges, surfaces, and shapes
- Pooling layers reduce spatial dimensions for computational efficiency
- Batch normalization ensures stability during training

Vision Transformer Backbone

- Pre-trained EfficientNetB0 backbone captures global contexts and hierarchical features
- Self-attention mechanisms model complex relationships within the image

Feature Concatenation

- Extracted features from CNN layers and Vision Transformer are concatenated
- Comprehensive feature representation combines local and global information for accurate classification

Model Training and Optimization Optimization Techniques

- Dropout regularization reduces overfitting by randomly dropping neurons during training
- Learning rate scheduling adjusts the learning rate dynamically for smooth convergence
- Validation checking monitors the model's accuracy and loss over epochs using a reserved validation dataset

Preprocessing Data Preparation

- Resizing: Images are resized to a fixed dimension (e.g., 128x128 pixels) for consistency
- Normalization: Pixel values are scaled to the range [0, 1] to reduce numerical errors and accelerate computations
- Augmentation: Techniques like rotation, flipping, zooming, and brightness adjustments are applied to diversify the dataset and enhance model robustness - Noise Reduction: Gaussian blur is used to diminish noise, ensuring better feature extraction and improving

detection accuracy

Implementation of Face Mask Detection Using CNN-ViT

To implement a face mask detection system using a hybrid CNN-ViT model, the system follows a structured workflow involving input handling, preprocessing, model training, and alert mechanisms. The approach ensures high accuracy and scalability for real-world applications.

System Components 1. Image Input

Images are acquired through:

- **Camera Feed**: Real-time image capture from connected cameras.
- **Uploaded Images**: User-provided images via a web interface.

2. Image Preprocessing

Preprocessing ensures consistency in input data:

- **Resizing**: All images are resized to a standard size (e.g., 128x128 pixels).
- **Normalization**: Pixel values are scaled to the range [0, 1] for uniform model processing.
- **Augmentation**: Techniques like rotation, flipping, and zooming are applied to enhance model robustness.

3. **CNN-ViT Model Architecture** ○ **CNN Component**:

Extracts local features like edges and textures using convolutional and pooling layers.

- **ViT Component**: Captures global feature relationships using a Vision Transformer backbone such as EfficientNetB0.
- **Feature Fusion**: Combines CNN-derived and ViT- derived features to ensure comprehensive analysis for binary classification.

4. scala**Classification**

The model classifies images into two categories: ○ **With Mask**: The individual is wearing a mask. ○ **Without Mask**: The individual is not wearing a mask.

5. Alert Mechanism

For mask violations, the system can trigger an email alert with:

○ **Subject**: "Mask Violation Detected" ○ **Message**: Description of the violation.

Attachment: Image showing the violation for verification.

Implementation Steps

1. Picture Input

Pictures are collected from real-time camera feeds or uploaded by users. These pictures are sent to the system for processing.

2. Preprocessing

The pictures undergo resizing to a fixed dimension and normalization to scale pixel values. Augmentation ensures better generalization during training by recreating different scenarios.

3. Model Training

A hybrid CNN-ViT architecture is designed to combine CNN's local feature extraction and ViT's global feature learning capabilities. The combined features are fed into fully connected layers for parallel classification. The model is trained on a dataset containing pictures of people with and without masks, split into training and validation sets.

4. Evaluation and Prediction

After training, the model evaluates input pictures, predicting whether they belong to the "with mask" or "without mask" category.

5. Mail Alert System

When a mask violation is detected, the system sends a real-time email notification to designated recipients with relevant details and the offending picture as an attachment.

RESULTS AND DISCUSSION

The real-world applications of the hybrid ViT-CNN model for face mask detection significantly rely on its inference speed. In this system, the model demonstrated low inference times, processing each image frame within milliseconds. This made the system highly suitable for real-time applications, allowing efficient monitoring of live security camera feeds or captured images without noticeable delays, thus ensuringtimely mask detection.

The model achieved high accuracy during training and validation, demonstrating strong performance in differentiating between masked and unmasked individuals. The training and validation curves showed consistent improvement over epochs, indicating effective model training. Additionally, the prediction system was capable of generating alerts for non-mask detections, which can be integrated into real-world monitoring setups.

However, certain challenges were identified during the development and deployment process. Lighting conditions proved to be a significant factor, as the model's accuracy dropped under poor lighting. Features like facial contours and mask boundaries were harder to detect in such scenarios. Similarly, environmental complexities, such as the presence of multiple individuals in crowded areas, led to occasional misclassifications.

Despite these challenges, the system demonstrated good scalability. By leveraging GPU optimization and batch processing, the performance remained consistent, even with higher resolution images or multiple camera feeds. The use of advanced pre-trained architectures like EfficientNetB0 ensured robustness and adaptability.

Overall, the system was satisfactory and scalable for practical applications. Future enhancements, such as the integration of advanced data augmentation techniques and adaptive learning strategies, can further improve accuracy under challenging conditions. This demonstrates the potential for deployment in realworld scenarios like public spaces and workplace environments.

FUTURE SCOPE

The hybrid ViT-CNN model for face mask detection demonstrates promising results, but there is ample scope for future enhancements and applications.

IMPROVED DETECTION ACCURACY:

To address the challenges posed by poor lighting and environmental complexities, advanced data augmentation techniques can be explored. Incorporating synthetic image generation, such as GANs (Generative Adversarial Networks), could help the model generalize better to diverse conditions. Using transfer learning with larger, more diverse datasets can

further improve detection accuracy across different demographic and environmental settings. Edge Device Deployment:

The model can be optimized for deployment on edge devices, such as surveillance cameras with integrated AI capabilities, to reduce reliance on centralized systems and ensure real-time detection even in areas with limited network connectivity. Multi-Class Classification:

Extending the system to detect additional attributes, such as improper mask-wearing (e.g., masks below the nose), no mask, or other face coverings, would enhance its applicability in public health scenarios. Integration with IoT:

The system can be integrated into IoT frameworks to create a comprehensive monitoring solution. For example, detected alerts could automatically trigger alarms, lock doors, or notify authorities via multiple channels such as SMS, push notifications, or dedicated apps.

ADAPTATION TO ADVANCED SCENARIOS:

Adapting the system for use in complex environments, such as crowded public spaces or areas with high movement, could include incorporating motion tracking and trajectory analysis to reduce false positives. Realtime heat mapping of detected violations across locations could provide actionable insights for better resource allocation in enforcement efforts.

Cross-Domain Applications:

Beyond mask detection, the model architecture can be extended to other domains, such as detecting safety equipment like helmets in industrial environments or identifying potential hazards in security setups. Cloud Integrationand Scalability:

Cloud-based solutions could enable seamless scalability for larger deployments, such as city-wide monitoring systems. Integration with

REFRENCES

I. 1.[Tan, M., & Le, Q. V. (2019). "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." arXiv preprint arXiv:1905.11946.

II. 2.Vaswani, A., et al. (2017). "Attention Is All You Need." Advances in Neural Information Processing Systems (NeurIPS).

III. 3.Chollet, F. (2017). "Xception: Deep Learning with Depthwise Separable Convolutions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1251-1258.

IV. 4.Keras Documentation. "ImageDataGenerator Class."

V. Available at: https://keras.io/api/preprocessing/image/

VI. 5.TensorFlow Documentation. "EfficientNet Model."

VII. Availableat:

VIII. https://www.tensorflow.org/api_docs/python/tf/keras/ app lications/EfficientNetB0

IX. 6.OpenCV Documentation. "Image Processing in Python." Available at: https://docs.opencv.org/

X. 7.Kingma, D. P., & Ba, J. (2014). "Adam: A Method for Stochastic Optimization." arXiv preprint arXiv:1412.6980.

XI. 8.Dataset Source: "Face Mask Dataset." Available at: [provide the dataset's link or citation]. web-based dashboards could offer real-time analytics, reporting, and user-friendly visualization of monitoring data. Sustainability Enhancements:

Researching energy-efficient model architectures and training methodologies can minimize the environmental impact of deploying the system on a large scale.

By addressing these areas, the system's robustness, scalability, and versatility can be significantly enhanced, making it a valuable tool across a range of real-world applications.

CONCLUSION

The hybrid ViT-CNN model for face mask detection effectively combines the strengths of convolutional neural networks for local feature extraction and vision transformers for global feature understanding. The system demonstrated high accuracy and low inference time, making it suitable for real-time applications, such as monitoring live camera feeds or processing captured images.

The implementation successfully addressed the primary objective of detecting masked and unmasked individuals and provided robust email notifications for real- time alerts. This feature ensures timely communication with stakeholders, enhancing the system's utility in practical scenarios.

Despite its strengths, challenges such as varying lighting conditions and environmental complexities highlighted areas for improvement. These limitations underline the need for advanced preprocessing techniques and larger datasets to improve detection accuracy under diverse conditions.

Overall, the project showcased the viability and scalability of the hybrid ViT CNN model for real-world applications. With future enhancements, such as edge deployment, multi-class classification, and IoT integration, this system has the potential to significantly contribute to public health safety and other securityrelated domains.