Improving Bangla Sign Language Detection from Thermal Imagery: Leveraging Thermal Heatmap-Induced Transfer Learning and Deep Neural Networks

by

Rakesh Rakshit
19101588
Mohammed Fackruddin Tusher
19301120
Mobashir Mahmud Faisal Moyen
19301116
MD. Rahik Tahmid
19301269

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Rakesh Rakshit 19101588	Mohammed Fackruddin Tusher 19301120
Mobashir Mahmud Faisal Moyen 19301116	MD. Rahik Tahmid 19301269

Approval

The thesis/project titled "Improving Bangla Sign Language Detection from Thermal Imagery: Leveraging Thermal Heatmap-Induced Transfer Learning and Deep Neural Networks" submitted by

1. Rakesh Rakshit(19101588)
2. Mohammed Fackruddin Tusher(19301120)
3. Mobashir Mahmud Faisal Moyen(19301116)
4. MD. Rahik Tahmid(19301269)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October 20, 2024.

Examining Committee:

Supervisor:
(Member)

Md Tanzim Reza
Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement (Optional)

This thesis has been conducted in adherence to ethical standards, ensuring respect, integrity, and transparency in all aspects of research. All data used has been sourced, processed, and analyzed responsibly, maintaining confidentiality and protecting the privacy of individuals wherever applicable. No harm, deception, or undue influence has been exercised during the research process. Furthermore, all findings and contributions have been presented honestly, without manipulation or misrepresentation, in order to contribute objectively to the academic and scientific community. I am committed to the ethical principles of accuracy, accountability, and respect for intellectual property throughout this work.

Abstract

This research explores fusion of deep learning for sign language recognition and thermal imaging, focused on advancing communication systems for the hearing impaired. Recognition of sign language is a crucial technology for improving accessibility and the use of thermal images light up new possibilities for gesture or sign recognition in different lighting conditions.To address the challenges associated with recognizing Bangla sign language, we have collected and built a novel thermal image dataset using the ABF Astron Infrared Camera F3.20. Additionally, we created 49 distinct classes for Bangla sign gestures using a colormap filtered version of the dataset to enhance feature visibility. The study employed transfer learning to retrain pre existing neural networks on both the colormap and thermal datasets. Three prominent deep learning models, ResNet, DenseNet, and Inception, were selected for this task due to their proven effectiveness in image classification tasks.These models were first trained on the colormap dataset. Subsequently, we tested the models on the original thermal dataset, using transfer learning to refine the learned features. This process showed the potential for improved gesture recognition when utilizing both thermal and color mapped images.The results highlight the importance of combining transfer learning with advanced neural networks to enhance recognition systems. With accuracies of 95%, 90%, and 98% across different models, the findings demonstrate the promising application of thermal imaging in improving the reliability and accessibility of sign language recognition technologies. This approach could offer more sturdy solutions for real-time sign language translation, particularly in challenging environmental conditions.

Keywords: Transfer learning; Colormap ; heatmap; thermal image ;bangla sign language; CNN;

Dedication (Optional)

This thesis is dedicated to our family and loved ones, whose unwavering support, encouragement, and love has been our constant source of strength. To our family, for their guidance and belief in us even during the toughest times; and to our friends, for their understanding and cheer through every step of this journey. This work would not have been possible without their enduring faith and encouragement. Thank you for standing by our side.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our supervisor Mr. Tanzim Reza sir for his kind support and advice in our work. He helped us whenever we needed help.
And finally to our parents without their throughout sup-port it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ϵ Epsilon

υ Upsilon

$BdSL$ Bangla dataset of sign language

CNN Convolutional Neural Network

HCI Human-Computer Interaction

IR Infrared

ODI One day International

RGB Red, Green, and Blue

Chapter 1

Introduction

1.1 Background work

The technology applied to recognize sign language has worn the red feather over interest, mainly because of systems capable through them to bridge a communication gap for people who are deaf or hard of hearing. One of these potential areas for application is BdSL (Bangla sign language C) gesture recognition. In fact, thanks to the deep learning realm and especially convolutional neural networks (CNNs) including some transfer learning methodologies which we will dig more soon on image-based gesture recognition has seen great improvements. Sign language is studied through different machine learning methods like Convolution Neural Networks, R-CNN and Xception to generate a real time recognition model of Bangla sign languages(BdSL). While efforts have been made in the past to progress over their work and recognize both static,dynamic gestures using models trained on different datasets [24][32]. However, it has demonstrated success using convolutional neural networks to identify hand signs in BdSL, and achieved high accuracy results for digit(recognize) as well as letter [24]. Even data scarcity has been overcome by using transfer learning, and advanced augmentation strategies to improve generalization in the model[32]. Therefore, we have proposed our research against "Thermal Imaging based BdSL recognition" that can add a new paradigm in the field by increasing feature extraction changes with thermal, colormap transformed images. This method is not only a dataset expansion but also helps in understanding how different image modality can affect the accuracy of gesture recognition, which could help in creating a more adaptable recognition system.

1.2 Problem Statement

Among the sign language recognition datasets, for both general and Bangla sign language are limited in number while being of insufficient diversity and scale. This limitation is severe, and it hinders the advance of building real-world perception or recognition systems which can generalize across different environments. Although thermal imaging has unique advantages for recording hand gestures in low light or varying conditions, its use as a means of sign language recognition is still relatively unexplored. This research aims to fill these gaps as well by constructing and using a dataset that combines traditional (RGB) sign language images with their thermal images. This study aims to increase the performance and reliability of state of the art sign language recognition systems using sophisticated deep learning models like ResNet, DenseNet, Inception-networks combined with transfer learning techniques. The purpose of this solution was to provide a recognition system that might prove more accessible and useful than previous ones, especially for the hearing impaired community by combining thermal imaging with color mapped representations so as to enhance visibility over a broad range of environments.

1.3 Research objective

This research focuses to develop an improved sign language recognition system which is suitable for Bangla sign language (BdSL) using thermal imaging, and color map fused datasets.

- We would make a dataset which includes thermal images of Bangla hand sign gestures and convert an already available BdSL dataset into colormap format This dual approach allows including multiple environmental conditions for example, temperatures, lighting and the gesture complexity that are lacking in existing datasets.
- We concentrate on training and fine-tuning the state of art deep learning models ResNet, DenseNet and Inception using transfer learning methods. These will be trained on the thermal or colormap datasets as well and enables us to assess how good they are at recognizing complex hand gestures in diverse conditions. Due to the differences in grammatical structure, etc,strong also helps fine tuning these models so they can better handle BdSL data which increases overall accuracy and validity.
- One of the main purposes is to investigate cross modality learning. How well can the proposed method perform when combining thermal imaging with colormap enhanced features? In this study, we compared two imaging techniques to enhance the performance of gesture recognition in demanding conditions like low light or temperature variations.

The research also aims to make strides towards giving hearing impaired individuals easier access by finding its place in the evolution of a more extensible and precise real time sign language recognition.

1.4 Hypothesis

This study intends to analyze the performance of Bangla sign language with thermal imaging integrated deep learning based models and expects higher accuracy, validity as compared to color based traditional datasets or conventional gray scale images. Leveraging thermal data allows for distinguishing heat signatures and hand movements uniquely under more diverse lighting conditions which we anticipate to result in better recognition performance. Furthermore, transfer learning on other models like ResNet and DenseNet and Inception will also benefit from the previous learnings to get a high accuracy in recognizing different sign language gestures which eventually delivers an open source tool for the hearing impaired community.

1.5 Significance of study

In this work, we propose a method for Bangla sign language recognition using thermal imaging that counters the limitations of RGB and grayscale systems. The study increases its accuracy, using deep learning models such as Resnet, DenseNet and Inception in challenging conditions that include low light or complex backgrounds based on the fusion of thermal data. Academically, our work contributes to the state of computer vision and multimodal data fusion, whereas it provides more affordable tools for communication in Bangladesh using hearing-impaired individuals as a test case scenario establishing groundwork towards assisting technology in HCI with future applications.

1.6 Scope and Limitations

In this research, it is aimed to design and assess a sign language recognition utilizing deep learning with thermal imaging which are specific for Bangla sign language (BdSL). This includes the development of 2 datasets: (i) RGB color-mapped Bangla sign language image dataset and ii) Thermal infrared images using an IR camera. The pre-trained models for training are ResNet, DenseNet and Inception, we have used transfer learning to achieve maximum accuracy. Our goal is to increase the recognition accuracy of sign language in challenging scenarios like low lighting conditions by incorporating thermal data. Still, there are a few caveats to bear in mind. Firstly, the size of our dataset is limited in comparison to that for other sign languages used for similar research questions and may limit how models generalize. Secondly, the study is performed on a monolingual dataset hence making it non-generalizable to an unseen sign language without retraining. Thirdly, while thermal imaging has benefits in low light conditions, it might not function as effectively in places where the temperatures can swing to extreme levels. In addition, training of these deep learning models may necessitate significant computational resources which preclude real-time deployment on low power devices. In short, this research presents a valuable method to enhance sign language recognition using thermal data and at the same time highlights its drawbacks regarding dataset size, coverage of languages and its real time feasibility thus suggesting areas for further work in future.

1.7 Structure of the Thesis

In order to do this, the analysis is divided into various chapters that will introduce step by step all aspects of our research and results. The opening chapter explains the introduction of background, problem statement and research objectives with rationale. The second chapter of related work recollects existing studies on sign language recognition including earlier efforts in regard to thermal imaging technology. The third chapter, the methodology is presented along with how data was gathered and preprocessed followed by several deep learning models used. Chapter Four presents the experimental results: accuracy metrics, confusion matrices and other evaluations. From there, the research wraps with a discussion of results and they study possible contributions and some potential future work. The discussion concludes with the limitations and practical implications of their research.

1.8 Thesis Statement

The results from current research show how thermal imaging of color-mapped Bangla sign language data can increase the performance accuracy using transfer learning architectures. We intend to show that when these datasets are combined, accuracy and reliability in sign language recognition increases, thus making communication technologies more easily accessible for the deaf. The results will help to convey the promise of state-of-the-art imaging methods in overcoming current limitations in recognition.

Chapter 2

Related Work

Recognition system proposed by Danier and Aveen [22] for robust hand gesture recognition via high resolution thermal imaging. This thermal imaging application, not the traditional RGB based systems where one cannot run through one's autonomous potter just because it is night and dark for a camera. In addition, the authors collected a dataset of 14,400 thermal hand gestures and attained classifying accuracy at approximately 98.81% using deep convolutional neural networks (CNN). The model was also tested on the edge computing devices namely, Raspberry Pi 4 and Nvidia Jetson AGX to show its deployment efficiency in real time applications where it achieved an inference time of around 75.138ms when run at NVDLA_convolutions_8bit.pt which executes DNN pruning from this repository. However, as Raspberry provides hardware support for single mode configurations only that is why actual optimizations are not reflected here as various Dual mode or other such power consumption settings couldn't be used so nowhere near the full implementation which I hope someone else can manage.This research's main concentration is on thermal image based gesture recognition but without any work done with respect to different environmental temperatures and occlusion conditions in real life that can lead to differing results in temperature readings. In addition, the exact gestures used to train their model and test it are also fixed leaving unanswered questions about how well this system would for general use with other hand movements or users. Additionally, even with the optimized image processing pipeline for this neural network inference system running effectively on OpenCL GPUs mentioned earlier, practical deployment of such systems in more resource constrained environments or detecting gestures that are far removed from a simple hand raise may still require strategies not yet addressed.

Researchers Daniel and Simen [21]come up with a new idea to recognize sign language digits from thermal images using machine learning. It looks at working around the limitations of color and depth cameras, especially in changing lighting conditions with thermal cameras. To classify thermal hand gesture images with robustness they proposed a lightweight deep learning model inspired from bottleneck structures in deep residual networks. The database has 3200 images with one digit inputs that are represented in low resolution by a thermal camera (32 x 32 pixels). The authors also added that one can also use a Raspberry Pi to integrate this model into an edge computing system, which is highly portable and cost effective. The model performed the best overall achieving an accuracy of 99.52%, showcasing effective use for thermal images in digit recognition with sign language. The study, however, has some shortcomings in that it boasts high accuracy and a compact edge computing system. A major limitation is the lack of data, with just 320 images per digit, which may prevent proper generalization on much larger and more diverse datasets. In addition, the thermal images are low resolution (32×32 pixels) so may not capture enough detail in complex gestures and especially beyond the digits. Another significant drawback is that the study only deals with sign language numbers, limiting its adaptable application for higher complexity sign languages gestures. Also, though thermal imaging can reduce lighting problems in these cases, there is still the potential risk of suffering from environmental heat interference or different hand sizes and shapes. Motivated from the limitations identified in this study, we extend a basic research on thermal imagery based sign language recognition. Taking into account the digits exploited in recent study, we extend this work by utilizing a larger dataset that covers more diversified Bangla sign language signs. Moreover, one can also utilize transfer learning mechanisms (from thermal heat maps) to make the model more generalizable for complex gestures.

Muhammed and Tonoy [33] tried to make use of the Convolutional Neural Network (CNN) approach, saying ConvNeural. This is achieved by two fine tuned datasets: "BdSL_OPSA22_STATIC1" and "BdSL_OPSA22_STATIC2" with 24,615 images and 8,437 images of Bangla characters and numerals respectively that helps to mitigate existing system´s limitation. For the first dataset, the ConvNeural model posted 98.38% accuracy and for the second it was around 92.78%, much higher than what pertained models were able to accumulate until now. The study was devoted to static images, which restricts its application in real time scenarios containing dynamic gestures despite the high accuracy. Additionally, the model is not tested on datasets with more varied backgrounds and lighting conditions. Moreover the study focuses on the numerals and alphabets but the dataset could be enlarged in terms of the different signs with some extended classes. Therefore, this drives us to conduct research that overcomes the constraints placed on static datasets, and helps in real time Bangla nonverbal communication recognition. And so, we expect to improve dynamic gesture recognition through the use of thermal imagery and then inducing transfer learning for more complex gestures by making it much suitable towards broader real world use cases.

In this study Xianwei and Yanqiong [5]tried to improve the SLR using CNN and RNN. The have used gloves sensors which can detect the movement of hand signs efficiently. In addition, some of the sensors used here are a leap motion Controller (LMC) that allows it to recognize hand and finger gestures. Beside, this LMC uses infrared sensors and cameras to track the finger gesture in 3D spaces. The work of LMC is basically tracking the movement of hand and figures in different angles and then interpreted it into sign languages. VGG net16 consists of 16 layers, additionally it has extra 13 layers which are convolutional and 3 layers which are fully connected .To collect data sets, the authors have used RGB cameras, depth sensors for better dataset quality . Nevertheless, this study needs a diverse dataset. Specific languages need specific datasets so lack of datasets creates the research user independency also a major limitation of these datasets.Through the researcher tried to solve real world problems but unfortunately sometime recognize it collapse and wasn't able to meet its target. On the other hands vgg16 is used here we know that, VGG16 is very slow and also it requires lots of disk space and bandwidth which is not efficient to train by a big datasets.Thats why we used vgg19 here to make the model faster and to make the accuracy higher, Another limitation is developing models for real time applications specially to use on other devices like mobile.

The authors Sai and Rakesh[31] propose a system for accurate hand gesture recognition using thermal imaging on edge computing, which is implemented based on deep CNN . It mitigates the defects of conventional RGB and depth cameras, which are ill suited to perform in environments with insufficient clarity or high complexities. The researchers first built a dataset consisting of 4,800 thermal images in ten hand gesture classes and then employed a lightweight deep convolutional neural network known as CNN model for classification. Their models were able to reach up to 99.79% and it worked even with stable results under different lighting conditions like direct sunlight, indoors daylight or in dark rooms. Furthermore, the study also compares inference time and accuracy of proposed CNN with pretrained models performing in a very computationally efficient way. Although the gestures in other tasks may not be so good, there are still some shortcomings. In addition, since the thermal dataset consists of only a small amount of data around 10 hand gesture classes, we do not know how well it would work for more complex gestures. Static hand gestures are responsible but it is important to note that the system lacks continuous gesture recognition, which would be necessary in sign language. Moreover, although the lightweight CNN we proposed is effective, high resolution thermal images could still be computationally expensive to process even with an efficient implementation of this model so that realtime applicability on low power or portable devices could remain difficult.

In this study , Aveen and Linga [28]propose adversarial unsupervised domain adaptation for hand gesture recognition in thermal images to tackle the dearth of labeled thermal datasets for hand gesture recognition. Therefore, deep learning has been employed as an effective solution to overcome the reliance on labeled thermal data, although it requires significant amounts of annotated images for training . The authors introduced an unsupervised domain adaptation known as UDA based on deep architecture that can be used without any supervision information from the fine tuning dataset or fully labeled target dataset. More concretely, they suggest a channel attention and bottleneck layers based adversarial UDA model for the task of learning domain stable features across RGB vs thermal domains. This enables the model to benefit from labeled RGB data for enhancing thermal image based hand gesture recognition. The performance of the model is demonstrated with respect to two datasets, sign digit classification and alphabet gesture classification, achieving an accuracy rate of 91.32%and 80.91outperforming existing methods in terms of accuracy,model size and parameter efficiency. Nonetheless, although the proposed model is successful in overcoming this restriction of insufficiently labeled thermal data, the model performs well on these two datasets but has not been tested for generalization capability to other hand gestures or real world scenarios. Furthermore, RGB datasets used as a domain adaptation source might not proliferate in all cases or they could come difficult with no integrated input. As well as the task of alphabet gesture classification losing a few points of classification accuracy with 80.91% rate, indicating further enclosure for improvement, such in more sophisticated gestures or larger databases supports this hypothesis on the model's performance limitations about external dataset and validation set quality. The computational requirements for adversarial UDA models may also restrict their use in resource limited settings.

First of all, The study of Deep an Chintan [25] proposed a model to recognize Indian sign language (ISL) gestures from video sequences and combining with LSTM and GRU to handle the vanishing gradient problem and they used recurrent neural networks known as RNNs. The goal of this model is to detect ISL signs and translate them into corresponding english words. Secondly, It splits a video which contains gesture sequences into smaller segments where each represents a distinct word. After that, these segments are divided into frames which are processed using Inception and ResNetV2 where the combination is a deep neural network that extracts key features. In addition,these feature vectors are passed to a multi-layer system combining LSTM and GRU units, which can remember prior inputs and capture long-term dependencies in gesture sequences. The model also uses dropout layers to prevent overfitting and softmax for classification. Thirdly, Experiments with six different LSTM-GRU combinations show that the LSTM-GRU configuration provides the highest accuracy. A custom dataset (IISL2020) of 16 subjects performing 11 ISL gestures was created for training and testing. Despite challenges like varying lighting conditions and non-ideal recording settings, the model achieved a high accuracy rate of 97% in recognizing signs. The research suggests further enhancements, such as expanding the dataset and addressing low-light conditions through image augmentation. This system is designed to work in real-world conditions, making it a step forward in ISL recognition technology.

The research of Shiva Narayana and Festus [23]is based on a communication between speech impaired people who cannot speak to non speech impaired people by recognizing British sign language(BSL). Here, the authors also built two types of models, namely convolutional neural network (CNN) and long short term memory known as LSTM . Comparatively , CNN > LSTM with 98.8% training accuracy and 97.4% testing accuracy with the tool matplotlib reargument where the CNN also had weighted precision and recall of 97% and 96%, respectively. Accuracy was atrocious for the LSTM, less than 50% accuracy. Consequently, the study is able to conclude CNN works better for BSL recognition. Although the CNN model obtained high accuracy, it was designed for only British sign language and not applicable to other sign languages or common gestures. The LSTM model did not perform well, and it shows some pattern aspect of the dataset is very difficult for the type of deep learning algorithm to recognize which may restrict its usage in continuous non verbal language. More importantly, the study has not been evaluated on scenarios of real world diversity in terms of background and lighting conditions which are crucial for generalizing beyond controlled environments. The diversity and volume of the dataset could also be increased to allow better generalization across more users, as well in other aspects.

The researchers Deepak and Sudhanshu [29] presented a CNN model that is trained on hand sign images containing English alphabets. The model was built using 26,000 grayscale images which includes the training set and validation set. The model is an architecture of a few convolutional and pooling layers with fully connected layer modules, significantly optimized using the Adam algorithm. The accuracy of the model on a dataset found in Kaggle was recorded at 96.7%. This study showcases the capabilities of power efficient hand sign recognition using deep learning in real time applications like assistance to individuals with motor impairments and virtual reality interactions. While the resulting model demonstrated a high degree of accuracy, it only recognized static signs for the English alphabet as opposed to continuous gestures or other sign languages. The grayscale images of the dataset may also not code enough complexity to match with real world scenarios, such as variable lighting or backgrounds. On the other hand, although this model is capable of real time use as suggested by its relatively fast runtime as compared to VGG16 and also mention regarding computational efficiency or ease of deployment in mobile or low power environments was made. To make it more practical, the dataset could probably be enriched in diversity and include dynamic gesture recognition.

The analysis of Yanqiong and Xianwei [2]runs a comprehensive study on the development and obstacles through using deep learning, considering technologies in visual communication language recognition (SLR). The authors discuss improvements in sign data acquisition, analysis and different neural networks such as CNN or RNN which have shown great potential for the recognition of isolated signs. Therefore, they also serve as an introductory model of sign language recognition and a significant research challenge in continuous sign language recognition known as CSLR, with this pushing the researchers toward sequential models such as TRes. While the above work shows progress, implementation issues such as dataset diversity and user independence still need to be addressed in order to improve the practicability of using SLR. Resolving these issues can enable deployment with enhancement for real time performance suitable for mobile applications. Optimization while deep learning has helped in advancing sign language recognition, the paper points out some limitations. Models are also struggling to interpret continuous sign language, which involves fluid transitions between signs. In all cases, the system fails to generalize across scenarios due to a lack of user diversity and context in datasets. In addition, the paper stresses a lightweight model suitable for mobile applications since currently no such machinery is under evaluated. Articulation where a gesture blends with another is a characteristic property of continuous sign language and also lack of articulation modeling guidance in most recognition systems. These are all challenges that reduce the practical SLR performance these days.

The paper "Static Hand Gesture Recognition in Sign Language" [27] suggests a deep learning architecture that is designed to unify CNN and various feature extraction . The images of hand gestures are preprocessed and background eliminated before being pushed through the three feature extraction streams. It takes the features and combines them into a final vector for classification. The model is more generally accurate, since it scores over a 99.8% accuracy on various sign language datasets and rotation including gesture ambiguity data exposing the incredibly robust and general nature of our proposed solution concept in solving this problem. The proposed model provides a solution worthy of addressing these new challenges, but it lacks the generalizability required to handle dynamic gestures due to its target on static hand gesture synthesis. Preprocessing steps such as background removal which are far from perfect in real time operation may incur poor latency characteristics of it that is not overriding with trained data. It has been tested on three datasets yet its generalization to more diverse, real world scenarios remains an open question. Also, if multiple feature extraction methods are used that could lead to high computational complexity making it difficult for use in lightweight applications or mobile devices.

In this analysis, Mahin and Farahnaz [15]proposed a method which is a CNN based model to recognize Bangla sign language(BSL) using CNN. This model is based on a public BSL dataset, retrieving skeletal features by classifying hand gestures from the images. It is able to recognize and categorize hand signs with 98.75% accuracy. This aims to assist communication of Bengalispeaking deaf and mute people and bridge the gap in building a robust BSL recognition system across computer vision, machine learning. This study is limited in that it only concerns static hand gestures, and cannot be applied to cases where continuous gesture recognition issues are of interest. The model's performance on real time is untested and as well with a complex background. Using a single set of data to train and evaluate the system could make it less generalizable to wider sign variations in Bangla or more diverse user populations.

The study of Pragya and Abdullah [13] is dealing with the efficiency of detecting BdSl in real time. They created a dataset named BdSLInfinite of 37 sign classes and a collection of about two thousand images. With this dataset, they trained a CNN model using Xception architecture with an accuracy of 98.93% and the average response time being about only 48.53 ms which makes it better than any other method in recognizing BdSL scores not just providing accurate results but also provides the best as well fast solution for real time detection. Nonetheless, the proposed model is only experimented on a dated set of 2,000 images which may reduce the scalability in case of more comprehensive real world scenarios. Besides, the analysis only considers static signs and does not consider continued or dynamic sign detection. The complexity of the background and differences between users were not studied which could affect its real time performance in an uncontrolled environment.

The exploration of Jahanghir and Sabbir [16] describes a proposed deep learning model to serve and mitigate the challenges of speech and hearing impaired people in Bangladesh through their sign language. The proposed model gives the Bangla signs language (BSL) alphabets using a distinctive CNN architecture and it was trained on the sharaLipi database. It is a system that wants to cover the communication gap between those people with any type of handicap and the rest. The model's success rate is 99.86% percent, which helps address the failings of previous research in BSL alphabet detection sector This current work shows how computer vision based solutions are increasingly equipped to translate sign language into a form more readily understood by the lay public. Equipped with high accuracy, the model is essentially static alphabet recognition and does not cover dynamic or continuous gestures. The study is based on one data set and even just 6 to 10 year olds from the BSL community are a diverse bunch. Additionally, the testing and if required adaptation of this test as per real environment environment is not enlightened that can alter practical applications using the proposed system.

The experimentation of Shahjalal and Rafiqul [35] proposes a novel application that identifies hand sign digits and converts spoken it into duplication as well as explicit form. The model has 92% accuracy rates in classifying hand sign digits and is designed to target individuals who have speech impairments. A TexttoSpeech engine and translator converts it to Bangla speech output. This system is designed to reduce social barriers and make communication more accessible for people with disabilities. The authors are excited about the potential impact that deep learning may have on communication deficiencies in individuals with speech impairment. Moreover, a 92% accuracy rate does not sound great because there will probably be real world situations where we need correctness and proper communication is paramount. More extensive testing on a wider variety of datasets as well different environments is required to investigate the feasibility in general everyday use.

The investigation of Hossen and Arun [7] describes a method for recognizing the static hand signs of the Bengali alphabet using deep convolutional neural networks which is known as DCNNs. Experimentations of this system have been done on 37 Bengali letters by having an image database containing 1147 images. The method used a pre trained network and trained only the top layers, achieving 96.33% on training data and 84.68% on validation data By doing this they hope to be able to improve the recognition of Bengali sign language and create a channel that more accurately transcribes gestures made in supporting communication for Bangladeshi deaf. On the other hand, the validation accuracy of the system, 84.68%, suggests that there may be some opportunity for improvement. The Dataset is a bit small so it might not be that generalized to large and divergent datasets or real world scenarios.

The fieldwork of Boshir and Dardina [19] presents a deep learning framework of hand movement recognition in Bangla sign language, which uses DCNN to extract features from 95 gestures. This framework will bridge the gap of communication between deaf People. For the new alphabets recognition system a dataset including 3219 images of six persons has been created containing 37 Bangla alphabets containing 8 vowels and 29 consonants. The proposed model had an accuracy of 99.22%, showing that recognition Bangla hand gestures with HSV and YCbCr color spaces for hand detection is performed elementary using the approach discussed on this introductory paper Conclusions In contrast, the dataset is a relatively small subset of 3219 images and may not be representative in realistic use cases where the data contains far more or heterogeneously captured information. There is no talk about the potency of this whole process if it has to be done in real time/dynamically.

Shahinur and Mahib's study [20]presents a CNN based method for recognizing and translating Bangla sign language (BdSL) characters in real time. This system facilitates the communication between deaf individuals and non sign language users. Therefore, the dataset for training consisted of 4600 hand sign images, collected using a HD webcam. They have achieved a validation accuracy of 99.57% and a validation loss of 0.56%, with the model successfully detecting all twenty six alphabets, ten digits in BdSL. In addition, the system learns continually with new images being captured in realtime on a regular basis. Lighting range, partially occlusion hand gesture and etc. are not considered in the system which is interesting from a research point of view as well. Likewise, scalability of the dataset has not been investigated.

The fieldwork of Kanchan and Muhammad [26] introduces a deep learning model employing transfer learning with ResNet18 to identify the Bangla sign language. The authors claimed that the model is able to interpret 37 static signs of the one handed Bangla sign alphabet also known as BSL and ISL. Therefore, they created a dataset consisting of 45958 images, where the fingertip patterns are represented in four color coded classes to enforce low intra class similarity. As a result, the model had great accuracy of 99.97% with sensitivity and specificity rates at a very high level of 99.49% for being able to provide real time assistance in ensuring effective communication and way forward using sign language for hearing or speech impaired individuals. On the contrary, they did not address variations in signing styles between individuals which may affect abstraction. Further, realtime performance in arbitrary changing environment conditions like variable lighting or backgrounds have not been well examined.

The study of Jahir and Moinul [34] introduce BengaliSign, a data driven machine learning based interpreting system for the language of hypersurface merry go round dedicated to resolve communication disparity among deaf and nonverbal people with hearing enabled population. By applying a convolutional neural network (CNN) with squeeze excitation functionality, it results in an excellent 99.86% classification precision in the KUBdSL dataset. To make the information accessible to more people, a smartphone application was also developed. Using Shapley additive explanations to explain the model's decisions. The researchers find that hard centric visual cues are important for sign prediction and human communication behavior. Constraints high accuracy of the system is mitigated by diversity in signing style and different users interpret signs. Additional tests are also needed for understanding the model performance under real world scenarios and environmental conditions along with various lighting and background.

The report [18] proposes a new model for Bangla Sign Language recognition and sentence production based on deep learning classification. It addresses the communication gap between people who rely on verbal communication and sign language users, where up to 5population have no use for spoken words. Proposed: The entire system has been structured by using Convolutional Neural Network (CNN) for training and recognition of Bangla signs which is further capable of providing an avenue through what the hearing speech impaired section may communicate. This research uses a multimodal system for better interaction and as an educational tool to learn Bangla Sign Language efficiently. The proposed system may suffer from a potential limitation for recognizing the regional signs and variations in Bangla Sign Language. There is also no comprehensive evaluation of the CNN model under various real world conditions (like different lighting and background). Additional validation is needed on a more comprehensive dataset to improve the generalization of this model across sign language contexts.

The study of Safayat and khandoker [11] presents a novel technique to recognize

Bangla sign language (BdSL) alphabet using convolutional neural networks. This study emphasizes the importance that sign language has opened to deaf and hard of hearing population aims in creating a system for Bangla sign alphabets recognized by which shape based detection, detecting alphabetic letters are accomplished by structural similarity hand signs. A CNN with multiple layers, that can recognize the hand sign of a dataset containing 50 groups of images correctly for up to 92% cases. It puts an end to the dearth of research in BdSL and also attempts at bridging a communication gap between deaf individuals and the general population. One of the limitations used a small data set which comprised 50 sets that could affect its generalization and robustness in diverse real world conditions. Secondly, even with 92% correct predictions may not be good enough in real world scenarios due to varying environments and signed by different singers. The dataset has the scope of extension to add more data and make it versatile enough for other environments, user variations etc.

The research of Aditya and Shantanu[6] proposed a deep learning application to recognize sign language with special attention due to communication gaps. The research explores the areas capturing static sign language gestures using color based information from RGB cameras through convolutional neural networks known as CNN popularly known as Inception v3. Image dataset preprocessing is well carried out and input quality improved, which allows validation accuracy to cross 90%. Specifically, the paper will present an updated survey on sign language detection by reviewing past works in this field and indicating some of these challenges for future research directions in critical technological applications. While 90% validation accuracy is a great result, these static gestures will certainly impede the model recognition of dynamic or even continuous sign language.In addition, the study does not provide enough information about how big and varied its data set is, which can affect how well the model will work for different signers or realworld signs. Future work is necessary for improving these limitations before it could be recognized correctly with more challenges and situations.

The analysis of G.Anantha and K.Shyamala [9] deals with the issues related to gesture recognition in Indian sign language using convolutional neural networks (CNN) over a novelly created dataset based on continuous selfies video mode. Recorded from five subjects across multiple scenes and viewing angles, the data set consists of 200 signs containing around 60 frames per sign seen by hearing impaired individuals. This paper investigates different CNN architectures and training data to improve recognition accuracy. Their model was able to achieve an accuracy of 92.88% and outperformed other classifier models on the same dataset thus showing that CNNs has been proved good enough for sign language detection algorithms as well. A promising recognition rate From our perspective, the study focused on only two five subjects and may thus not provide a comprehensive account of sign language users. The focus on Indian Sign Language restricts generalisability with other sign languages, without considering the possibility of different signing styles. Likewise, the model must be further validated in real life scenarios to confirm its practicality but also reliability due to potential variations as a result of effects such as lighting and background noise.

The research of Lee and Goh [10] aims to identify static gestures of the American sign language (ASL) using convolutional neural networks (CNN). The primary motivation behind this was narrowing the communication void between deaf-mute individuals and common people. This study trains and validates the CNN model with a dataset of 4800 images. This expresses a high inter-class recognition accuracy of the model which is 95% with resilient ASL patterns containing total static images of 24 classes. The success of this CNN model offers an option for future, more sophisticated machine sign language translation systems that could improve the lives of people requiring flexible communication modes like deaf and hard-of-hearing individuals. Despite the high recognition accuracy of this model, it is only applicable to static ASL gestures and does not reflect the rich variety common in sign language. The dataset, however large, might not generalize well to the variation in hand shapes and movements of different users. Moreover, its practical utility in real-world conditions and for continuous signing or dynamic gestures has not been confirmed yet which could restrict the model's potential deployment across a wide range of scenarios.

The exploration of Devjoyti and Abdulla [14] has developed a model which is aimed at recognizing Bengali sign language. The research effort ensured addressing the existing and significant challenge of effective communication with and among the deaf and hard of hearing populations. Precisely , the authors used convolutional neural networks to compile the dataset Bengali Ishara-Lipi including a total of 6,760 images, 5,760 of which were allocated for training and 1,440 served to test the model. The overall recognition accuracy rate of the model was 92.7% for Bengali character sings. In this regard, the presented approach is aimed at contributing to the eventual creation of commercial device interpreters, therefore enabling effective communication between sign language users and the general population. However, while the model is accurate in recognizing signs, it has a range of limitations, such as based on a fixed dataset and limited to static signs, that prevent its immediate real-life implementation. In addition, the generalizability of the OBM has not been tested on a varying range of users with different singing styles, and the general population for applicability on a broad scale. While significant progress has been made for Bangla gesture Language (BdSL) recognition as shown in the reviewed papers, most of them deal with static signs and use simplistic CNN models generally obtaining mediocre accuracies. Now although these models for example: ResNet18, Xception tend to deliver great results but many times the model isn't robust enough for dynamic applications or they do not seek advanced architectures which can capture multiple complex features and generalize well. That inspired us to look further for more advanced models, such as DenseNet,Resnet and Inception. These residual connection models, known for capturing fine-grained spatial information DenseNet, exhibiting deeper layers for improved feature learning ResNet and multi-scale features extraction Inception, account for a higher accuracy robustness in determining the delicate thermal sign language clues present between thermal and visual modality.

The study of Pukar and Trilochan [12] presents the first self-powered thermal imprinted flexible triboelectric sensors known as TIFS that have been developed for SLT. This sensor can simulate a structure similar to human skin and use the contact electrophysiological effect and so it is PVDF-based with high sensitivity (0.5 V kPa [1]) and has a response time of fewer than 20 ms. Therefore, a wireless SLT system was successfully completed by integrating fourteen TIFSs into an MCU showing its feasibility during speech synthesis both alphabetically or wording through machine language from hand signs to voice motions for reading Text-to-Voice correspond on cell-phone screen. However, an issue with such a system is that it has to depend on the 14 sensors for sign language translation which may limit scaling up and likewise poses higher complexities in capturing an extensive compendium of signs. Moreover, even if the response time and sensitivity are praiseworthy those results might have to be better validated in different real world conditions such as from environmental factors or long term use sensor like behavior or accuracy could probably be compromised.

In This research paper [17]proposes different methods for hand gesture recognition using computer vision.As usually we know that hand gesture recognition can be used in different field like HCI,sign language interpretation, robotics etc.Hand gesture recognition can be two types such as static and dynamic.Also technological approaches can be used which is also described here like glove based and Computer vision based sensors.Techniques and challenges are also described here are skin color detection,appearance based recognition and motion based recognition,skeleton based recognition and depth based recognition. The challenges that are faced here are lighting variation,background complexities ,real time processing in diverse environments .In my paper I have also faced the same kind of limitations and I tried to overcome those though some problems are still there but I used Bengali language which was not done previously.

In this paper [1] author built a system for sign language recognition by 3-D hopfield neural networks.The technical site used here is first of all it matches the hand movement using Hausdroff distance and a measure compared the hand shape.Then major frames selected through video and compared the frames by hopfield neural networks with the model data.These are the technique discussed here.So the system tested with 15 different types of hand gesture and gained 87% accuracy.So it worked overall very well the limitation faced are complex background and creation larger database needed further study according to this paper.

Like others this paper [30] also introduced a system about thermal video based hand gesture recognition system using lightweight CNN.This system also identify hand gestures but it uses lightweight CNN for developing that system that is the uniqueness of this research paper.traditional gesture system uses RGB cameras which has many problems like lighting issues,skin issues etc.To overcome those problems and to achieving high accuracy author used CNN model to establish that system.Author proposed thermal images to overcome limitations of RGB cameras For tis author used FLIR Lepton 2.5 thermal camera.The datasets have 9 different hand positions with multi framing of thermal images.the proposed model used some hidden layer including max pooling layers,dense layers and dropout layers and batch normalizations.The accuracy achieve here was 97%,96% and 87% for the datasets of 97%,96%,87% for datasets of 15,10 and 5 frames Author also wants to improve this model by exploring ResNet architecture.Our proposed model uses ResNet DenseNet and inception like architecture for Bengali language models

Chapter 3

Dataset Description

3.1 Bangla dataset:

This dataset contains Bangla sign language images, more specifically, synthesized taking into account previous datasets which tend to be low on diversity and quantity of the data needed for developing a real life sign language recognition system.

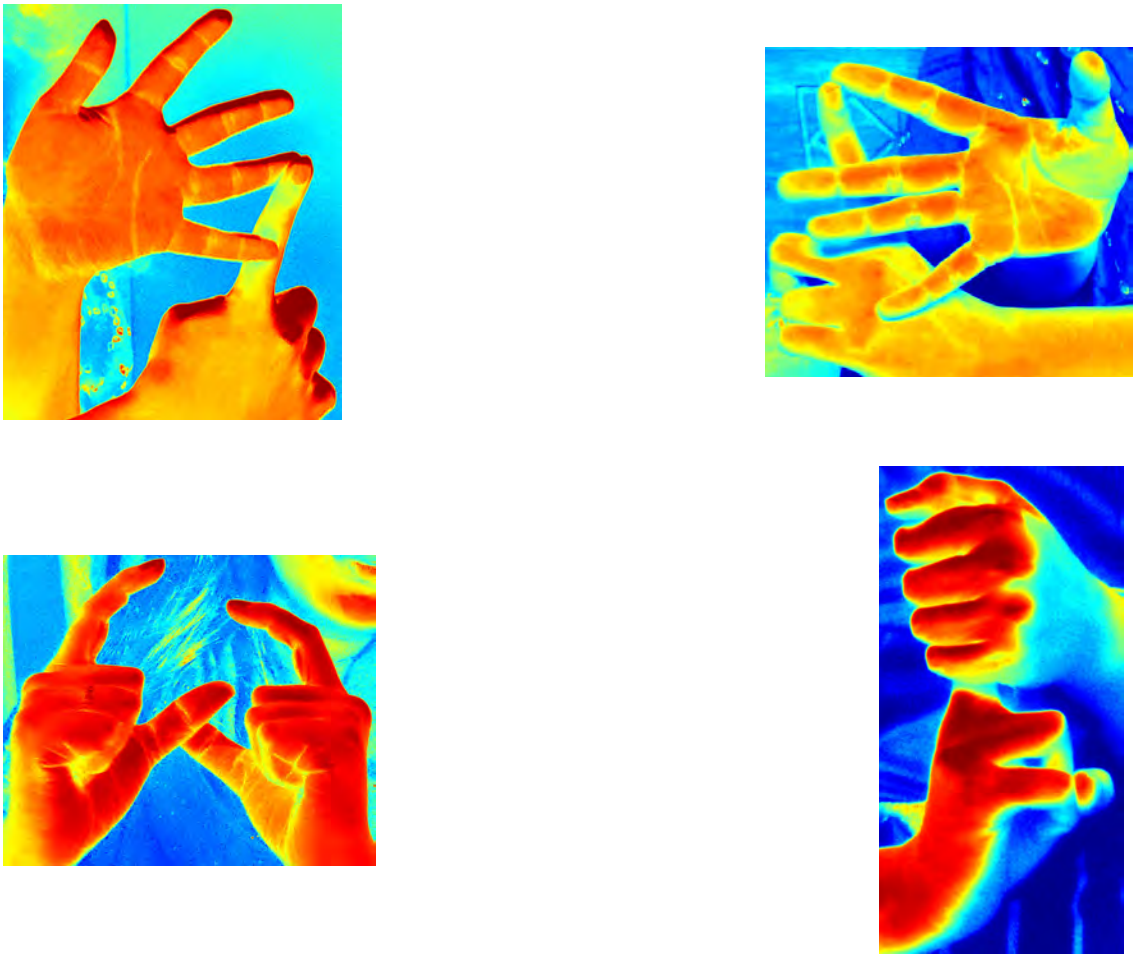


Figure 3.1: Data sample of Bangla Dataset

The first part of the dataset is intended for detection model training and consists of full-frame images with annotations. In the second part, which is used in our studies of recognition, models appear cut off without background hands signs only. Both sections contain images captured under a wide range of environmental conditions, such as varied settings in the background and lighting scenarios to increase manifestness across different environments. There were 14,714 images in each section that included a combined total of 29,428 among its respective composition for the same set filled by only up to overall measurable limit categories equaling out at around 49 per part. We chose the second part of this data set in our work, which was taken from a collection stored on Mendeley Data. We followed it up with a colormap filter on all the images to extract visual features in order for them to be processed and trained downstream.

3.2 Thermal dataset:

Bangla Sign Language Thermal Image Dataset was developed to inspire and assist researchers working with sign language recognition in thermal image sensing.

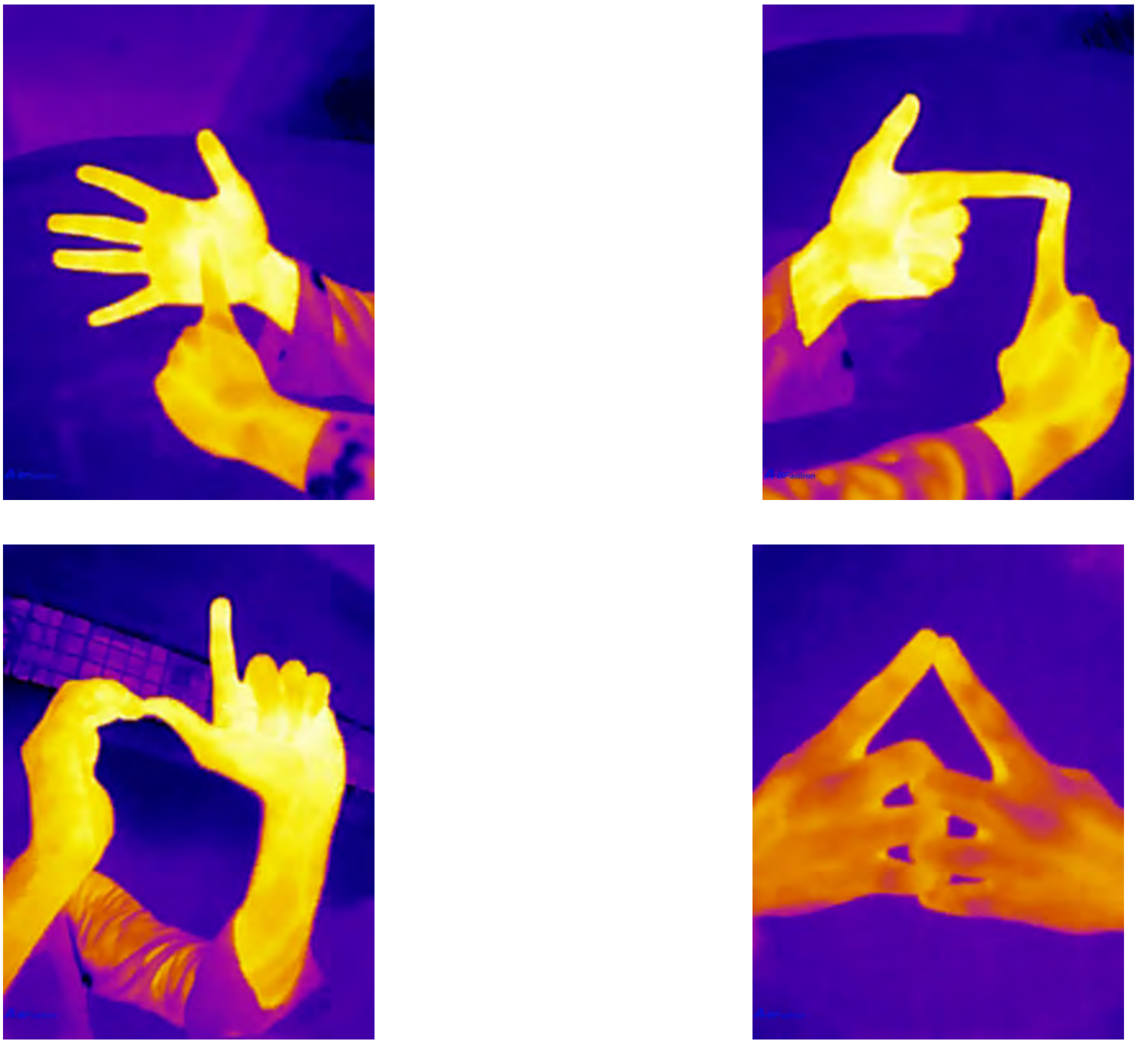


Figure 3.2: Data sample of Thermal Dataset

The dataset consists of 5,118 thermal images categorized into 49 classes corresponding to different Bangla sign language gestures. The images were obtained with the (ABF Astiron Infrared Camera F3) connected to a smartphone via USB Type-C, allowing for real-time thermal imaging. There were 17 males and 15 females, aged between the ages of 19 to 45 years, participating in total data collection. The samples are collected in deep cold and opposite hot climates that bring you temperature differences and noises. The shockwave function will prompt the dataset code to download an unfamiliar series of images each measuring at (160 x 120) pixels. The dataset's 20% sample was allocated for the test, with the remainder used for training to allow sturdy model evaluation. All annotations were done manually according to the accompanied sign language gesture specified within a given parameter for consistency. Yet, we acknowledge that the dataset available has some restrictions like possible biases in gesture representation and environmental conditions at image capture.

3.2.1 Camera details:

(ABF Astiron Infrared Camera F3 The DRS watch duty 20) is an application specific thermal imaging suite used for surveillance, industrial inspection and scientific research work. It is a (160 x 120) pixel resolution camera allowing us to see where temperatures differ. Normally, it sports a high resolution that will let us detect tiny changes in temperature for superior image quality. The camera has great thermal sensitivity measured as noise equivalent temperature difference (NETD), a very important feature for the resolution of minor temperature variations. It can work in a large temperature range of-20°C to 400 °C (-4°F to 752°F) and provide an accuracy near ± 2.0 C°: However, without the conventional camera display, rechargeable battery or even built-in storage . Instead, it connects to smartphones through a USB type-C port and makes use of the display and storage capabilities of such phones using a specialized app. In addition, because of this novel design in the sense one can now view and record thermal images in real-time, it will actually change how much use people extract from them. With a high thermal resolution and the capability for detecting temperature variances with precision, the camera is able to work in all relevant temperature ranges providing key solutions for various applications where CIGS DLB monitoring of temperature distribution is needed.

Chapter 4

Methodology

Our research uses deep learning for sign language recognition by considering thermal imaging and color mapped Bangla sign hand gesture datasets. At first we gathered and preprocessed two sets of data . First one, is a set translated from the existing Bangla sign language image dataset to color coded images, as well as another one is collected through an infrared camera A-BF astiron. A few preprocessing steps were applied in training such as resizing the image, normalization, and data augmentation to enhance generalization of the model. Analogous transfer learning was later performed with deep learning models from ResNet to DenseNet and Inception. This is the process of experiments we use to train models on a color mapped dataset, before fine tuning on thermal data for improved performance. Models were evaluated by classification accuracy, confusion matrix and ROC curve. Finally, the top performing models were saved, and examined for how these methods could be practically implemented in the future to help potentially improve accessibility hearing-impaired.

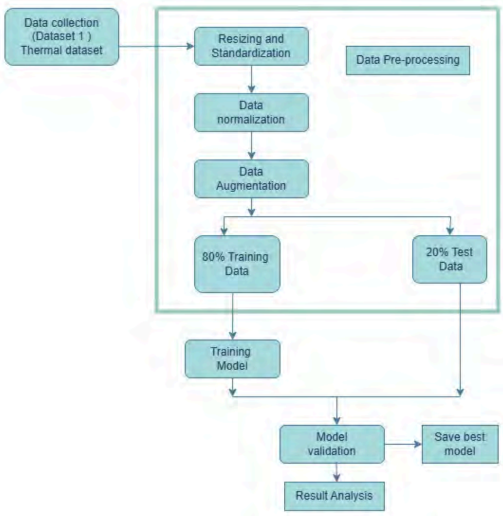


Figure 4.1: Thermal dataset training flowchart

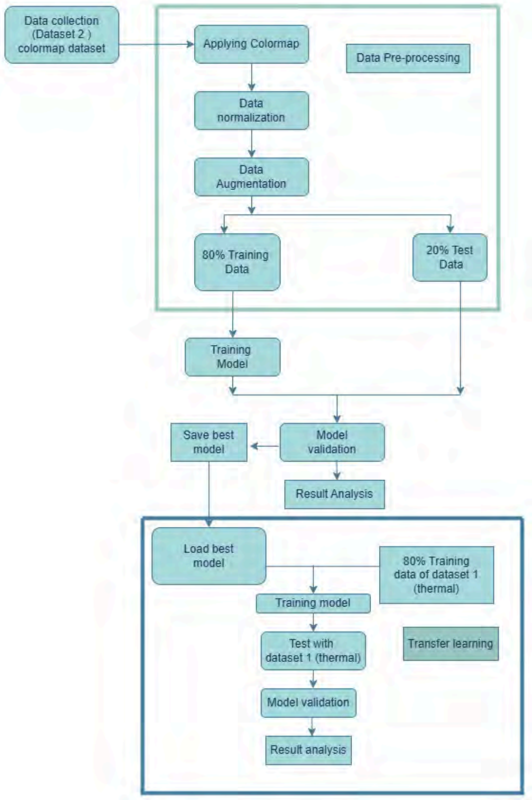


Figure 4.2: Transfer learning flowchart

4.1 Data Pre processing

For both the images and thermal Bangla sign language data, a preprocessing pipeline was defined respectively to convert the image representations of motions during signing such as colormap converted one and normalized red entail representation into computerized elements for deep learning model training. There were the core steps which made everything consistent and generally improved model performance, but also a specific step was there that only was applied to this dataset namely applying colormap filter.

4.1.1 Resizing and Standardization:

Both datasets were resized to 224x224 pixels as required by input for deep learning models such as ResNet, DenseNet and Inception. Training the models requires that all images have a consistent size, and we used resizing to ensure this.

4.1.2 Normalization:

The pixel values of both datasets were normalized by rescaling them to [0, 1] range using the ImageDataGenerator(rescale=1./255) function. Applying normalization was important to scale down pixel intensities and standardize them within the same level so that ultimately it helped faster convergence of training algorithms.

4.1.3 Data Augmentation:

Data augmentation was performed on both datasets to reduce the likelihood of overfitting and to improve generalization capabilities for our model. This involved random transitions such as rotation, horizontal flips, increasing brightness,zooming in or out and width or height shifts, otherwise posing to represent the closest real world scenarios of a human hand moving. The additional augmentations had the dual benefits of allowing us to train in a more diverse way and helping our models generalize better across signing conditions, such as different positions or orientations from varied signs.

4.1.4 Label Encoding:

The class labels in both datasets were converted to numerical format using Label encoding for models based on deep learning. It was important because the datasets have 49 classes of unique Bangla sign languages and so we wanted to prepare our data in a model-friendly format for classification tasks.

4.1.5 Colormap Filter:

Additionally, the colormap dataset passed through an extra processing step where a colormap filter was introduced to turn gray images into pseudo-color representations. This transformation, performed with OpenCV's cv2. The applyColorMap() function with options such as COLORMAP_JET, helped enhance the visual properties of the images identifying critical features making it relatively easier for models to differentiate between different signs.

4.1.6 Train Test Split:

Finally, it was divided into train_test_split with an 80-20% ratio. This allowed us to train the models on one subset of the dataset, while testing them on an unseen part of it and thus meaningfully assess their ability at generalizing. These preprocessing steps unitedly prepared the datasets ready for training deep learning models, which played a backbone to train a Bangla sign language recognition system that is not only accurate but also resilient. The only main difference was that we utilized the colormap filter in the dataset and began with a representation of improved features from images before training.

4.2 Model Description

4.2.1 Resnet50:

ResNet50 is a type of deep convolutional neural network which is practical on the vanishing gradient problem generally encountered in very deep networks[4]. It is done with the idea of residual learning. The second method involves skip connections to allow the gradients flow better during backpropagation, thereby allowing the network to learn direct mapping of identity functions. ResNet50 solves this issue of performance degradation as with the increasing depth, and ResNet50 enables training at much deeper models without loss in accuracy. Since then it has been demonstrated to be very powerful in image classification tasks and thus is a strong candidate for the case such as sign language recognition.

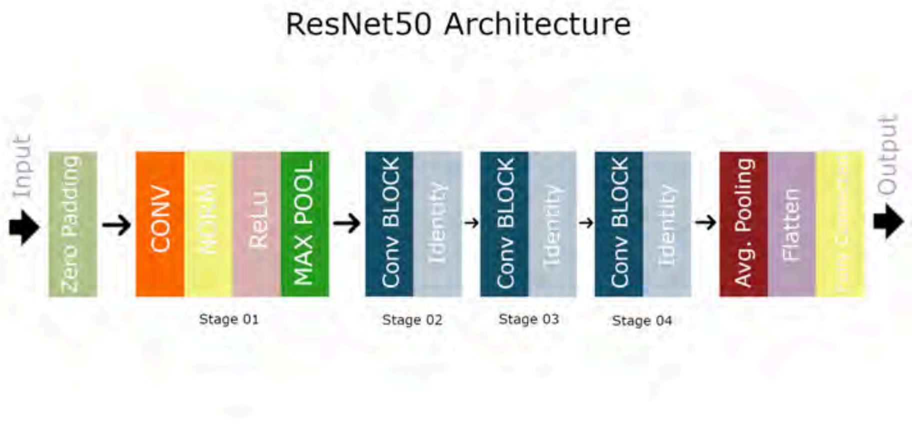


Figure 4.3: ResNet50 Architecture

In the above code, we used a ResNet50 architecture with 50 layers that are grouped by stages. The first step in the initial phase of operation involves a 7x7 convolution layer training with soup64 filters, then immediately followed by max-pooling to significantly reduce the spatial dimensions of input images. This is then followed by 5 stages in which the network uses both convolutional block and identity block together. While the projection shortcut is used to match input and output dimensions in convolutional blocks, the identity block retains the same dimensionality so that skip connection can act as a simple addition of layers. Every block has 2 or more convolutional layers with most using the smaller 3x3 size. What makes the ResNet50 model special is that it uses residual blocks in which input will be added to output of convolutional layers, a concept used similar with highway connection allowing networks to learn residual rather than relearning full transformation. The reason this design is unique is to avoid any problems with very deep networks like vanishing gradients. In addition, this model adds batch normalization and activation layers after each convolutional layer to make the training process faster and prevent our hidden units from saturating. ResNet50 uses bottle design which helps in enhancing the computational efficiency. It instead uses 1x1 convolutions at the beginning and end of each residual block to reduce and then restore the number of channels, which ultimately lowers the computational load by passing less data through all layers. This architecture makes the model faster in predicting, as well as requires less memory making it suitable for large scale applications. This is relevant to our investigation because of ResNet50's demonstrated competence with complicated visual tasks like those needed for identifying hand gestures in sign language. Deep residual learning, for example, purportedly can learn intricate image details and generalize better to unseen data. Therefore, ResNet50 architecture with deep structure and skip connections leads to abstraction of subtle differences in signs, which results in a major role improving the sign language recognition system. On the whole, ResNet50 remains a fundamental and dominant tool in computer vision landscape transforming image classification and identification association challenges crosswise.

4.2.2 Inception V3

Inception V3 is a convolutional neural network architecture [3] from the family of Inceptions, which were introduced in and originally developed for the ImageNet large visual recognition challenge.

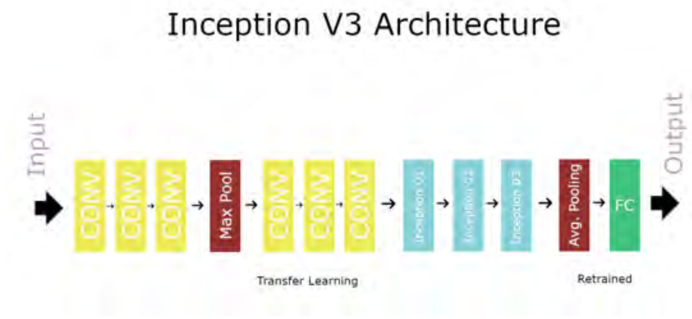


Figure 4.4: Inception v3 architecture

The inception module in Inception V3 is a very unique concept used in the network, it takes the usage of multiple filter size parallel processing inside a single layer. This architecture helps in capturing features at multiple scales, which enables the model to detect more complex patterns and shapes within images. Inception V3 uses 1x1, 3x3 and finally the larger convolutional filters increasing richness of expressiveness without significantly affecting computational cost.(4) Inception V3 has been constructed using the inception module series stacked as a topology, representing hierarchical feature extraction processes deep while being efficient. One of the good improvements in Inception v3 is to add factorization on some smaller convolutions. A similar example is that a 5x5 convolution can be separated as two cascaded 3x3 convolutions without losing the ability to extract complex features but reducing computational complexity. Inception V3 also uses asymmetric convolutions, 1x3 and 3x1 convolution to improve the range of features that can be learned by network .(4) A notable feature of Inception v3 is its use of batch normalization to normalize the convolutional layer outputs, which speeds up training and increases stability. During training, the model also uses auxiliary classifiers , which are additional layers attached to intermediate layers and provide extra gradients that help alleviate the vanishing-gradient problem. It helps learn better features in the lower layers ensuring that every region proposed has enough information to make a good prediction, hence helping it perform well during final classifications.(4) With the fact that we are working with complex datasets like those related to sign language recognition, something versatile and strong was needed. Inception V3 has more variety in comparison to ResNet50. It is possible to understand complex patterns present in the gestures with this architecture while keeping computational efficiency making it suitable for real-time applications. It is applicable to many practical uses due to the generic task performing level in various tasks, which has arisen as a popular selection for computer vision mainly and had an impression on gestures recognition systems.

4.2.3 DenseNet121

DenseNet121 (Densely Connected Convolutional Networks) introduces a new architecture that concerns connectivity of the layers[8], this enhances feature reusability while alleviating vanishing gradient problem. Unlike traditional convnets where each layer is connected to only the previous and subsequent layers, DenseNet121 connects every single one of them. This dense connectivity pattern helps the flow of gradients in backpropagation, and also facilitates learning more fine-grained features at different levels of abstraction.

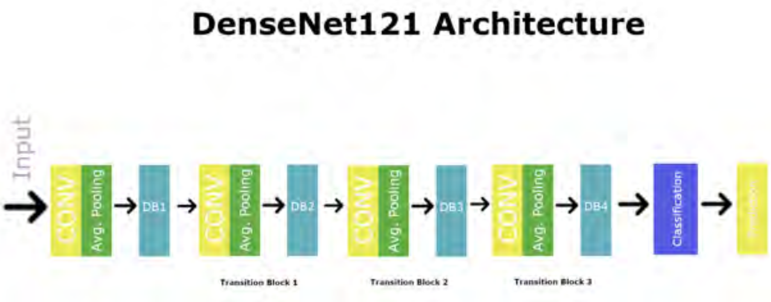


Figure 4.5: DenseNet architecture

DenseNet121 architecture consists of 121 layers grouped into dense blocks. There are many blocks, each block is a cascade of convolutional layers and every layer in the same block does receive input from all previous layers. The output of each layer is concatenated with the outputs from all previous layers, allowing reuse features and decreasing the number of parameters required in a network. This architecture is very novel and reduces the overall complexity of the model while retaining a lot of accuracy. Bottleneck layers are also used in DenseNet121 to decrease the dimensionality of input feature maps prior to convolution. By reducing the need for expensive dense connections, this method is demonstrated to help reduce computational burden while preserving important characteristics. Also, the compression factor between dense blocks is applied that reduces a number of feature maps and hence more efficient computationally. One of the key benefits provided by DenseNet121 is its ability to deliver high level precision with fewer parameters when you compare it against deep networks like Resnet and Inception. This efficiency is ideal for resource-constrained environments, making DenseNet121 an excellent choice in real world sign language recognition applications. Its capability to learn rich feature representations help it differentiate between slight variations of the hand gestures, an important part for accurate recognition. We chose DenseNet121 to build the model for our research as it has demonstrated state of the art performance on image classification and generalization across different datasets. With its original architecture it promotes learning fast and representing features well, providing a strong baseline for pushing the state of art in sign language recognition to provide better accessibility for hearing-impaired people.

4.2.4 Transfer Learning

Transfer learning is a machine learning approach which enables us to start with pre-built models to be developed for one task and then reuse them as initial starting points when modeling newer but relevant tasks. The motivation behind this is to take advantage of what the pretrained model has learned usually from big data, for example ImageNet in image classification. While the models are not trained from scratch, using these pretrained models gives them access to feature extraction capabilities that have been learned. In practice, this is referred to as transfer learning and simply taking the convolutional base of an already trained network and adding new dense layers that are going to be learned for your actual task. This fine tuning could be whole model training or top layers only, freezing the earlier layer to adapt your new data without relearn for features from a pretrained network. A consequence of this is that using a contrastive loss reduces the amount of data and computation which is needed to train a performant model. Transfer learning has become one of the most powerful tools in machine learning, with applications spanning image recognition with natural language processing as well as reinforcement learning . Taken together, this approach allows us to train models efficiently by transferring existing knowledge leading to superior downstream performance and reduced costs in terms of time or data required for learning new tasks.

Chapter 5

Result and analysis

5.1 Performance of Original Thermal Dataset

The model accuracy results show that different deep learning models used in this study have a distinct performance difference. According to the given dataset inception resulted with highest 92% accurate, making it very good neural network for recognizing complex hand gestures in different conditions. Similarly, ResNet achieved 90% accuracy which also showed its resilience towards the usage of two datasets; thermal and colormap. DenseNet is also performed well, however DenseNet just achieved an accuracy of 81%, that means it may not capture the complex negated features Bangla sign language gestures in relative manner with the other two models presumably . These results show that Inception and ResNet are the best algorithms for this task, followed by DenseNet.

Confusion matrices: The ResNet50 confusion matrix shows that it has a strong diagonal and this means the model is able to predict most of the classes correctly.

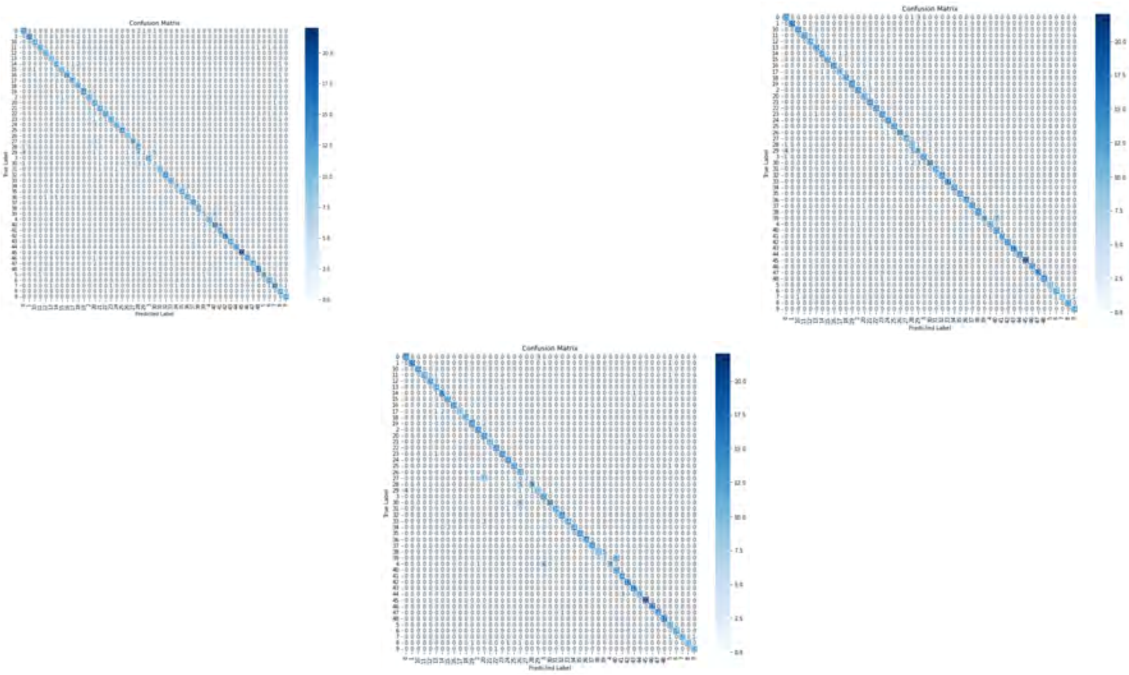
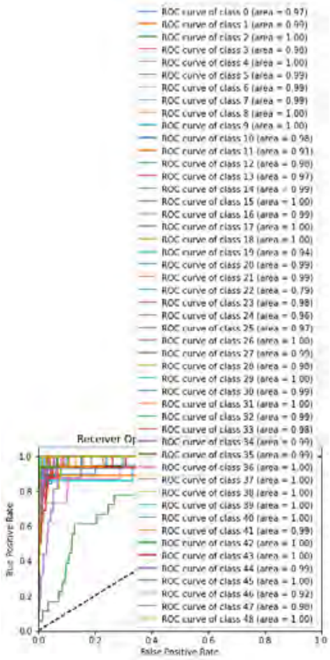
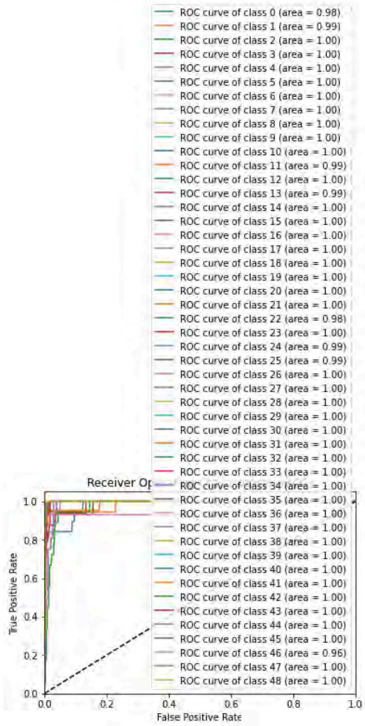


Figure 5.1: Confusion matrices for DenseNet, Inception, and ResNet thermal dataset

A spread of a very few misclassifications are scattered throughout the diagonal. However, few misclassifications present but are isolated whenever they arise. It means that resnet50 is good in general but maybe not as good for some classes.
As a result, the highest accuracy of all classes on most classes shows the potential of ResNet50 in distinguishing among different classes with minimum errors. In contrast, classes which are visually similar or share common features that may cause misclassification. The inceptionV3 confusion matrix also looks like there are more off-diagonal elements compared to ResNet50. This suggests that while InceptionV3 is high in general, it has a relatively high amount of misclassified data. Nevertheless, the off-diagonal elements are more sparse which indicates that the classes have higher confusion and so InceptionV3 could indicate it has more difficulty separating some class pairs. It still does a good job of not classifying most samples wrong. There is more mislabeling than in ResNet50, suggesting that it may fail to distinguish certain features or samples near class borders. DenseNet121 shows a strong diagonal indicating good classification performance. However, the density of off-diagonal elements looks slap in between ResNet50 and InceptionV3. There are still classification errors and the model appears to make fewer mistakes compared to InceptionV3 but likely more than ResNet50. They do seem to be a little clustered in places of misclassifications. Better than InceptionV3, with less mis considerations. On the contrary, some pairs of classes are still problematic, although not as much as InceptionV3 since fewer misclassifications seem to happen again and again.
Roc curves: Looking at ROC curves when benchmarking ResNet50, InceptionV3 and DenseNet121 various conclusions can be drawn.
ROC curve (Receiver Operating Characteristic): A graphical representation that shows the trade-off between True Positive Rate (TPR) and False Positive Rate(FPR) for every classification threshold applied. A key metric extracted from the ROC curve is AUC (Area Under the Curve), which represents how well a model can differentiate between classes. A perfect model would have an AUC value of 1.00, and a value of 0.5 represents no discriminative ability better than random guessing The performance of ResNet-50 has been demonstrated in the ROC curve, with AUC values close to 1.00 for all classes except one class at higher range but not so much high as VGG16 and InceptionV3 which is >0.990 only a few are less <0.950 or lower (Fig systemFontOfSize2). The model shows clear classification, more than 90% of the classes are perfectly separable. For a couple of the classes like class 1,11 and 24 it goes little down around to be 0.98–0.99 AUC which still is pretty much perfect model reached so far ResNet50 is the most consistent model across classes, with a very small difference in its performance. The model's performance of distinguishing between the hot and or other thermal images is dramatically better, as can be seen from the ROC curve. On the other hand, InceptionV3 also performs really well with over 1.00 AUC for a lot of classes. There is a little bit nuance in performance of few classes like class 0,1,2 and24 where AUC around 0.97 to 0.99 due its rarity Although these numbers still indicate great classification performance, they imply that InceptionV3 finds some thermal image classes which are a little harder to distinguish than ResNet50. Although InceptionV3 has a slight disadvantage, it is still an excellent model and displays significant results with almost perfect precision values for all of the other classes. The last one (DenseNet121) has a good but less stable performance compared to the other two models. Varying AUC values are observed between classes, and while most classes have high inputs close to 1.00,





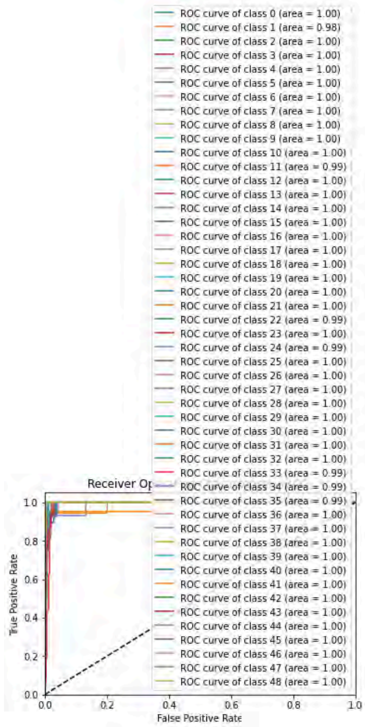


Figure 5.2: ROC curve of DenseNet, Inception, and ResNet on thermal dataset

there a few sharp drops in some of the class. For instance, class 22 and 38 have the lowest AUC values of about 0.79 and 0.76 implying that DenseNet121 has difficulty making distinction among these classes in stronger terms than other embeddings are able to demonstrate by means above. Furthermore,classes 1,12 and 37 have AUC scores of between 0.94 to 0.97 which is again indicating a performance drop with ResNet50 as well InceptionV3 These inconsistencies indicate that DenseNet121 might not be able to deal with the challenges presented by complexity of some image classes in thermal images which may make its classification model less robust on a whole. Finally, ROC curves and AUC values demonstrate that ResNet50 has comparatively the most stable performance across all classes out of other models with high metric value (Table 2), which makes it as model more reliable for thermal image classification. InceptionV3 is the next best, with somewhat different AUC values for some classes. Quality of DenseNet121, though still good in general but some classes show more unreliability and might be less effective - especially for real applications. Based on the comparison, we can conclude that overall ResNet50 performs better for this dataset.

Detailed performance metrics: Different performance patterns can be observed in the metric wise classification reports of three type models(ResNet, DenseNet and Inception) on thermal dataset w.r.t precision, recall and F1-score(IOException).

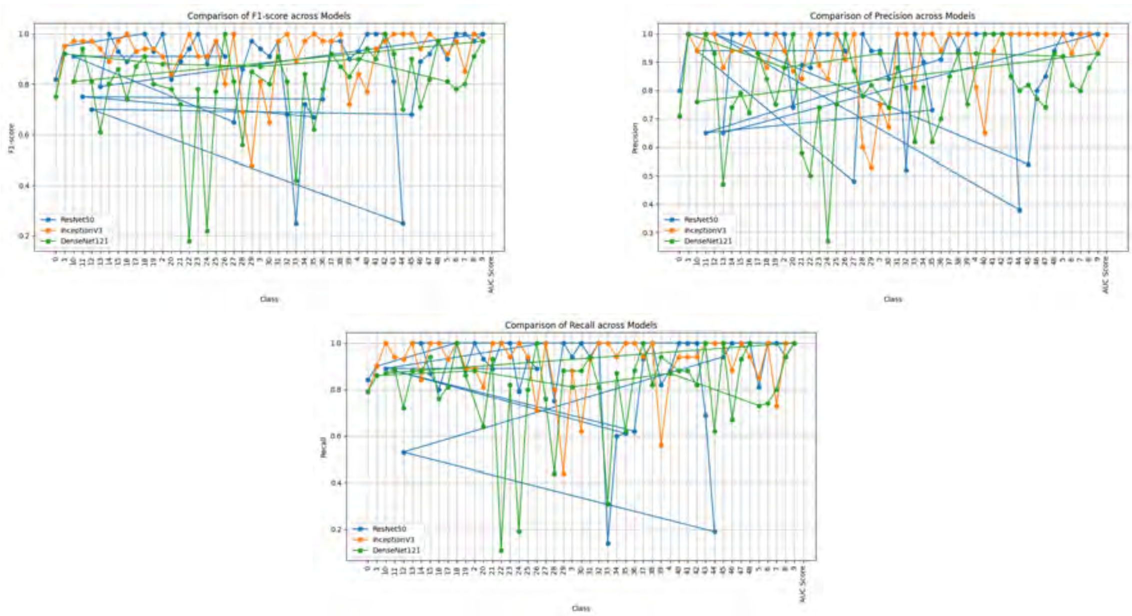


Figure 5.3: Comparison of F1-score, Precision and Recall across models of thermal dataset

Inception always displays a strong presence in class-wise metrics, as it outperforms ResNet and DenseNet 0–4 your model yielded an accuracy of only 91% with AUC being at most 0.9967 due to confusion between classes '2' vs '3'. However it achieves both higher precision as well raise and a better recall in almost all the classes, makingit robust to be able classify correctly the instances while minimizing false positives & negatives. Among the best performance, Inception achieves a perfect F1-score of 1.00 on many classes such as class 10,16,22,35–38 and from 45 to48 indicate high prediction score it got through all strong predicting power than any other model in this work ResNet also give good predictions which has an 88% accuracy and score of AUC as.9979 Austin slightly higher than Inceptions. It achieves a very high precision and recall for some classes, like class 1-10-16 or a correct one when real density in images of the Steele RUS settings shown above, Fig. It specifically struggles with class 27, the recall of which drops much more to 0.14 and affects the F1-score a lot. Nevertheless, ResNet remains competitive even for some classes, but again not as uniformly good as Inception. On the other hand, DenseNet has a comparatively less performance of 81% and AUC score of 0.9813 than ResNet and Inception. Especially in difficult classes, it exhibits the lowest precisions for scores of precision, recall and F1-scores. For example, in class 29, DenseNet achieves a recall of only 0.11 which had an obvious negative impact on the final performance. Nonetheless, DenseNet has good absolute performance on some classes (it obtains a perfect recall of 1.00 in class 17), however its general pattern is more unstable than the other two models To sum up, all three models perform well on thermal dataset but inception turned out to be best model which provides the high precision asleasrabe score and F1 scores for most of classes. ResNet is quite close by (high IoU, tightly distributed scores), and DenseNet lags behind a bit more particularly in the harder classes.

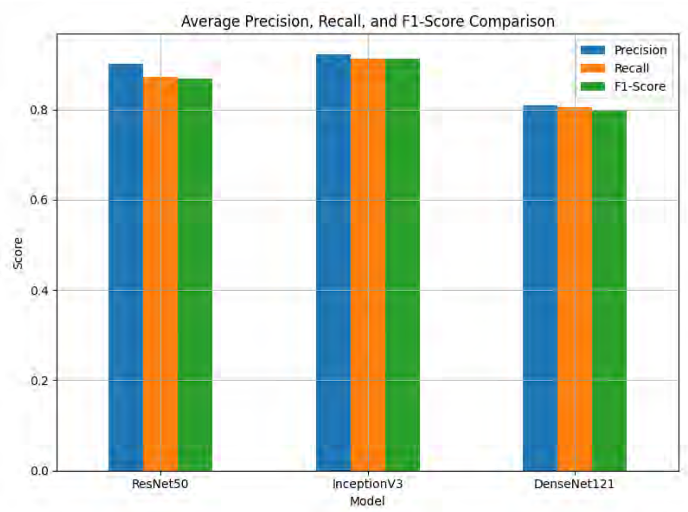


Figure 5.4: Average Precision, Recall and F1- Score Comparison on thermal dataset

5.2 Performance of colormap Dataset

Further analysis, showcased how DenseNet 121 achieved a performance of 81%, InceptionV3 reached upto an accuracy rate of 84% and ResNet50 outperformed them both with the classification success level increasing to as much as 98%. ResNet50 has the highest performance since it uses a novel residual learning mechanism with skip connections to address vanishing gradient problem, therefore allowing us to build even more deeper and efficient systems capable of capturing complex features. The depth of this network, with the ability to retain essential feature information through all layers makes ResNet50 highly suitable for handling complex data such as colormap. This is in stark contrast to InceptionV3, which has an accuracy of 84% due primarily to the fact that it processes multiple filter sizes at once using a unique architecture and so can capture features from different scales. But it is not as efficient in handling deep hierarchical features like what resNet50 would do and its accuracy on the dataset will be lower. DenseNet121 has 81% accuracy and uses a dense connectivity structure, which encourages feature reuse. But in large datasets it can result in feature redundancy and its simpler, relative to ResNet architecture doesn't allow capturing fine details on the same level. This is also the reason why ResNet50 can preserve information through holistic learning and manages to stay on top outperforming other models.

Confusion matrices: From the above confusion matrix of Resnet50 on color map dataset, we can see a clear diagonal pattern in the figure which means ResNet50 did well and predicted most no of hand signs to be correct.

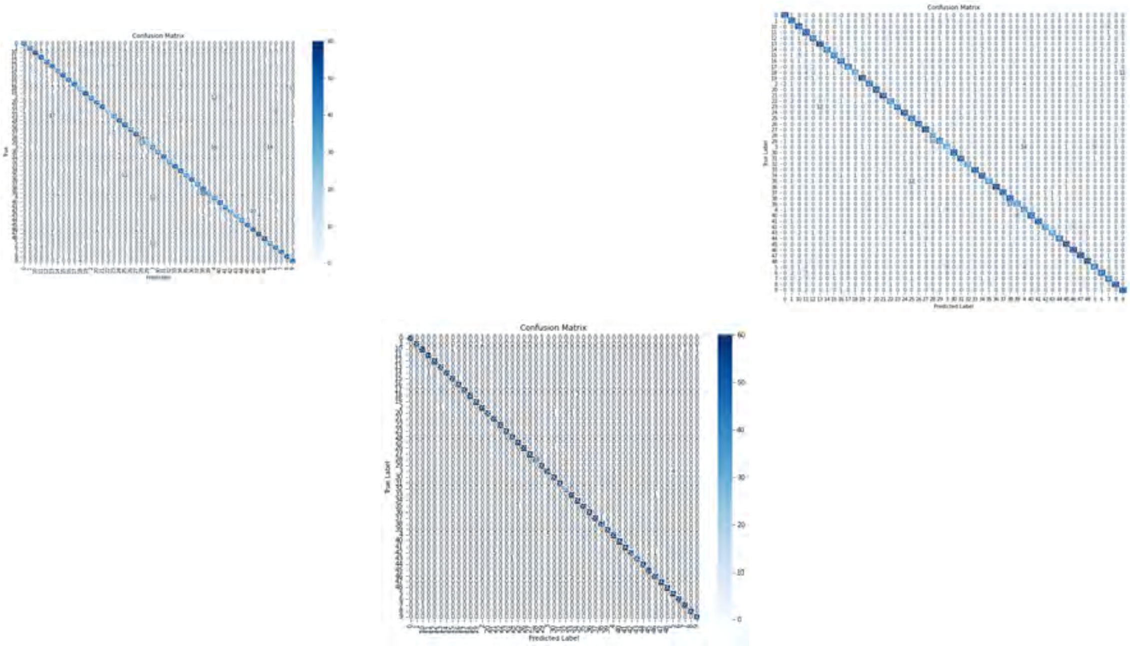


Figure 5.5: Confusion matrices for DenseNet, Inception, and ResNet on colormap dataset

There are misclassified off-diagonal elements in the matrix which suggest that one class be confused with another. However, some classes have been misclassified, which is evident in the white cells outside of the diagonal. The errors could be results of the similar hand gestural aspects that were needed for those specific signs. To conclude, the count of correct classifications is high and it also provides few off diagonal errors indicating that ResNet50 does a great job capturing data with good accuracy. InceptionV3 shows the same clear diagonal pattern like ResNet50, hence good performance overall. Nonetheless, there seems to be a wider spread of incorrect classifications as compared to ResNet50 with some clear errors in specific regions. No notable patterns are noticed here, and also we have to note that the (x, y) axis labels are not in order of ascending classes of misclassifications. Relatively few more misclassifications than with ResNet50, which likely indicates that InceptionV3 struggles to differentiate some character pairs from this ordered set. In general, InceptionV3 did not perform significantly worse than with ResNet50 in terms of the strictness accuracy level, However using slightly extended paths for predictions lead to a flatter off diagonal misclassification showing that it might generalize only moderately better on some other hand sign classes compared to what can be obtained from performing approximately as well but always making across class errors. DenseNet121 has the most pronounced and focused diagonal pattern that demonstrates high classification performance among these three models. There are only a couple of off-diagonal errors. Nevertheless, there are some misclassifications with DenseNet121. The majority of predictions reside on the 1:1 diagonal implying that this model outperforms in learning some relevant features to distinguish between various hand classes. DenseNet121 seems to produce the fewest misclassifications among the models. On the confusion matrix, it shows that this process could identify more hand signs with higher precision but less mistakes.

Roc curves:

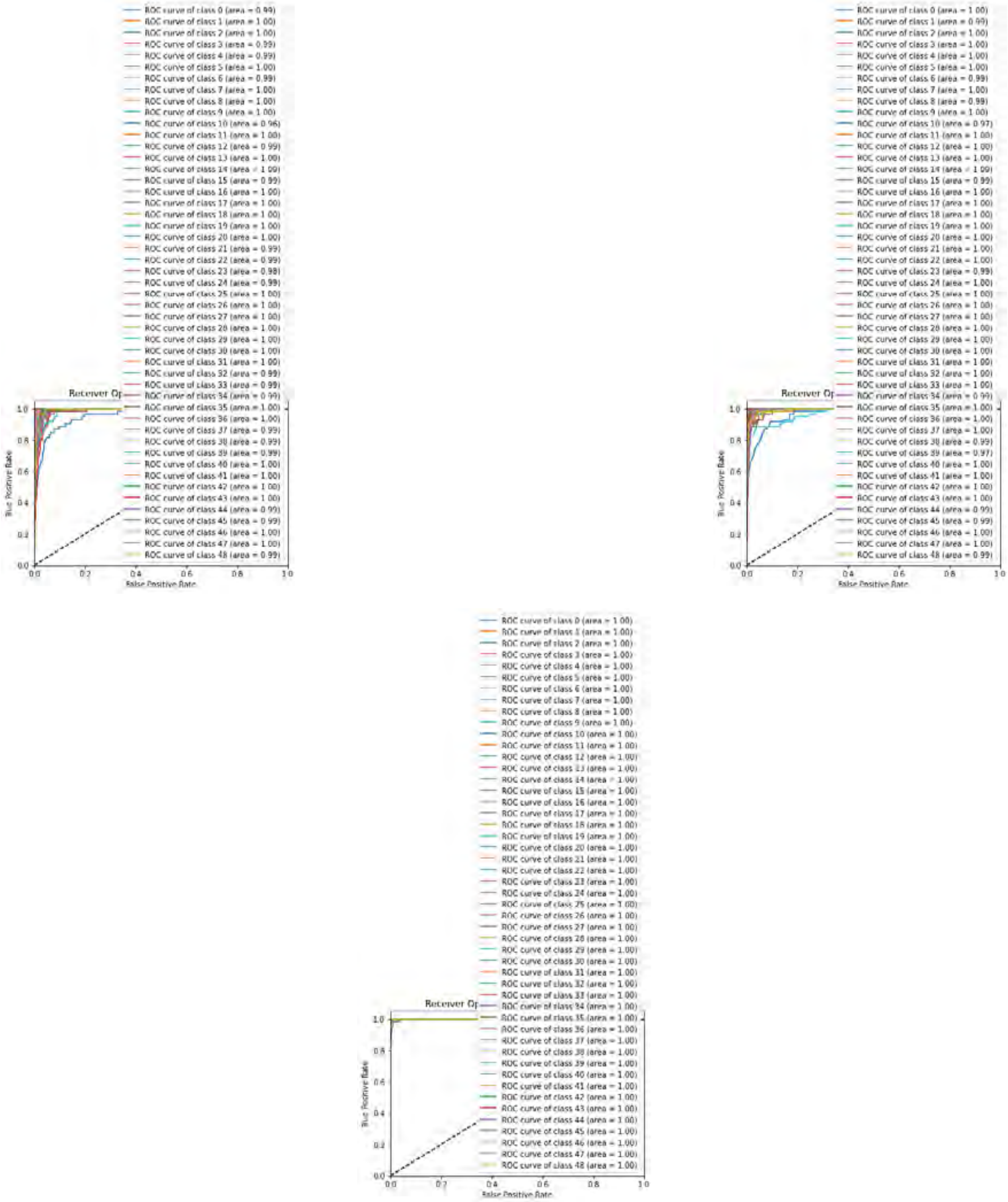


Figure 5.6: ROC curve of DenseNet, Inception, and ResNet on colormap dataset

DenseNet121, InceptionV3, and ResNet50 all attained excellent performance on the ROC curves for a multi-class target. The ROC curve of DenseNet121 is pretty smooth for almost all 49 classes and the AUC scores for most are approaching 1.00. Thus, DenseNet121 worked really good on classifying classes clearly with less False Positive at very high True Positives. It is the dense connections in its architecture that make it so efficient at capturing these complex patterns and thus allows SOTA performance across almost all classes. Some classes see only modest decreases in AUC, falling to 0.99 on each of them showing almost no challenges at all for classification. InceptionV3 also shows good classification capacity but it lacks a bit in terms of consistency when compared to DenseNet121. Although some classes start to have lesser AUC values (around 0.97 — class10,38), most of the ROC curves are still perfect up until a threshold crossover point These minor differences indicate that InceptionV3 has slighter difficulties making clear class boundaries in certain cases, which perhaps is a property of its architecture: whilst able to pick up heterogeneous features quite well (recall the small standard deviations), it might not consistently succeed with more subtle class separations as DenseNet121 or ResNet50. Yet, its results overall are still very strong — achieve most of ROC curves clumped in the top-left corner. ResNet50 also matches performance with Dense [8](for instance their ROC curves closely cluster at the high top left corner where perfect classification for all classes would be observed). Though the plot does not show it (because of my stupid font size), almost all classes are performed with an AUC score that is very close to 1.00, which means this network effectively suppress false positive rate and have high true positive rates on each class implementations. Even though there are small drops in AUC but they only reduced slighty and they have almost no overall effect on the performance. Given the deep layers and shortcuts in ResNet50, it tends to generalize relatively well across all classes with similar consistency as DenseNet121 but somewhat outperforming inceptionV3 at defining fine-grained class distinctions.
Detailed performance metrics: Considering the performance of ResNet50, Densetnet121 and InceptionV3 in relation to each other on colormap dataset shows unique strengths for each model.
Results: ResNet50 always outperforms the other models whereas DenseNet121 and InceptionV3 had variable classification capabilities across different class classes. Overall accuracy of ResNet50: 98% AUC Score :0.999 It was able to perform with great success across almost all classes and it reached nearly perfect precision, recall and f1-scores for most of the categories including ones that were class 0, 1,12,19,22,36.47.48 It achieved an f1-score of up to! ResNet50 has been used here because it worked very well and there seems to be a minimum error rate with the highest precision & recall values, which shows that even if we face data imbalance issue for these signage classes at such high-levels of accuracy ResNet50 is capable enough differentiate between all those different classes. ResNet: This paper demonstrated that a deep architecture of more than 100 layers outperforms the state-of-the art solution because it can learn complex patterns in an optimized way and eliminates overfitting configurations. It was most notably better at separating out classes with little overlap which made it the best performing model of our three. Though a strong performer, DenseNet121 was marginally the lowest in accuracy with 80% and an AUC score of almost exactly one point lower than ResNet at around.995 (indicating it has struggled more on some classes compared to ResNet). Though a

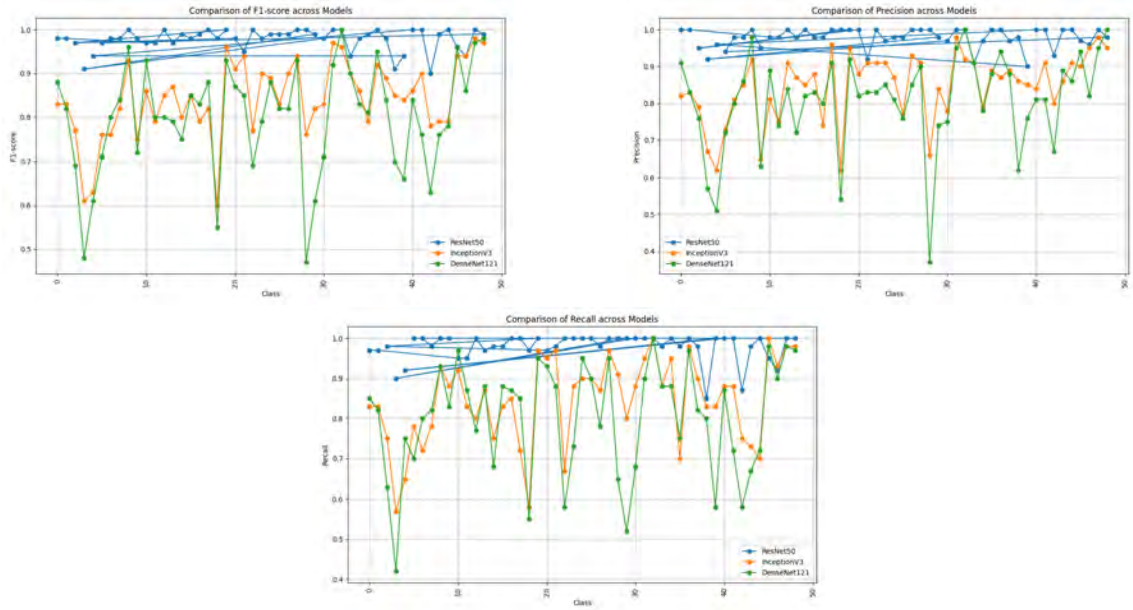


Figure 5.7: Comparison of F1-score, Precision and Recall across models of colormap dataset

model like DenseNet (Table 7) was able to obtain an impressive f1 score of around 0.93–0.94 in some categories such as class 19 and in class 27, the performance also suffered notable precision recall drops for example on classes where it gave only 55% average f1-score for all images under each event slide). The fact that existing works [13, 12] show such variation in performance across classes indicates possible difficulty of the model to perform well when features are less clear or more ambiguous and may partially explain incorrect layer selection despite efficient reuse. Nevertheless for most classes the breadth could manage f1-scores of approximately 0.80 to 0.90 but its inconsistency against restnet points at a small gap when compared in robustness.

Meanwhile, InceptionV3 had an accuracy of 84% with a AUC score of about 0.996 (better than DenseNet but also worse than ResNet) InceptionV3 had a strong performance for high-precision classes — class 31 (0.97) and 47 (0.98), both achieved near-perfect f1-scores It is interesting to note, InceptionV3 was the best performer on low-prevalence groups as well: 6 It did have trouble with some classes such as class 18 and class 3 (0.60, and f1-score of 0.61 respectively), showing the difficulty in accurately identifying more complex or noisy classes however InceptionV3 did pretty well in terms of providing a sweet spot between precision and recall for these classes, but failed sometimes when class overlap was greater or the differences were more subtle. Despite an overall better performance than the DenseNet, its inconsistency in some categories still brought it down to be unable to match up and surpass ResNet's classification.

ResNet50 was the standout, as it had great AUCs for each class consistently and near perfect value. It has shown the superior ability to discriminate between classes and produced most accurate classification results. Although still competent, DenseNet121 was more inconsistent with certain classifications causing significant levels of precision and recall to drop on specific classes. InceptionV3 performed (as expected) between those two models and without getting to the outstanding performance that

ResNet has in these situations — especially when harder or class overlap comes into play.
In all, as shown below image ResNet50 edges out most of the defeated DenseNet111 and InceptionV3 emerges between which is again a decent edge also DenseNet121 does well still over here shares more inconsistencies/ varies it performance.

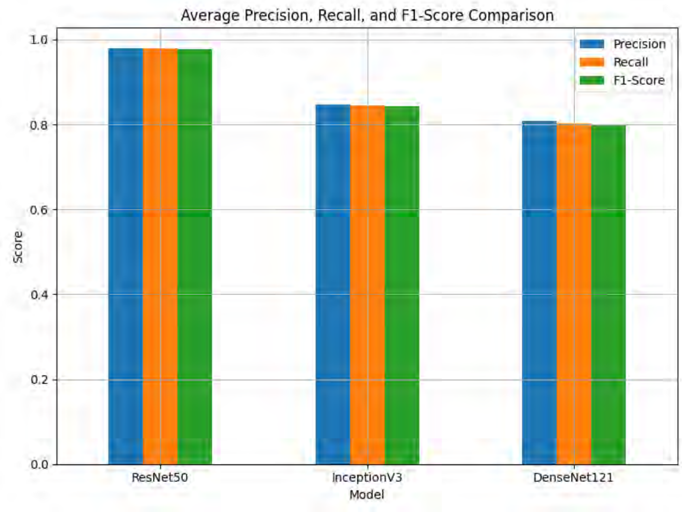


Figure 5.8: Average Precision, Recall and F1- Score Comparison on colormap dataset

5.3 Performance of transfer learning:

Transfer learning accuracy: yields an unbelievable 98% precision regarding transfer learning when InceptionV3 outperforms the other models in their capacity to use pre-trained weights for new tasks, retaining high levels of readability. Like-ResNet50 is not any less, keeping the performance bar really high with 95% accuracy on validation set due its deep architecture and residual connections which avoids vanishing gradients problems and enhances learning in deeper layers. DenseNet121 (90%) — Dense connectivity, no real data augmentation done this only dropout/lr iterations. In general, the three models are getting good results for transfer learning where InceptionV3 dominates in accuracy.
Confusion matrix: The diagonal in this confusion matrix of Resnet50 transfer learning seems to be quite strong, indicating a good performance with correct classification through most of the 49 classes.
Even so, the lone classification errors are occasional and closely surrounding the same labels with values that neighbors each other. Therefore, this matrix represents nice classification accuracy with low off diagonal noise. One good classification performance of the InceptionV3 confusion matrix of transfer learning is that the diagonal, as like the ResNet50 matrix, is also clear and dominant. Although, There are slightly more non-diagonal faulty classifications compared to ResNet50, nonetheless these still stay on the low side. Moreover, It seems that the errors are distributed a bit off by class when compared to ResNet50, but it is slightly more spread across classes which would mean InceptionV3 may classify not as good as on this dataset. We can see from the confusion matrix of DenseNet121 through transfer learning that it exhibited an extremely strong diagonal, even more coherent than

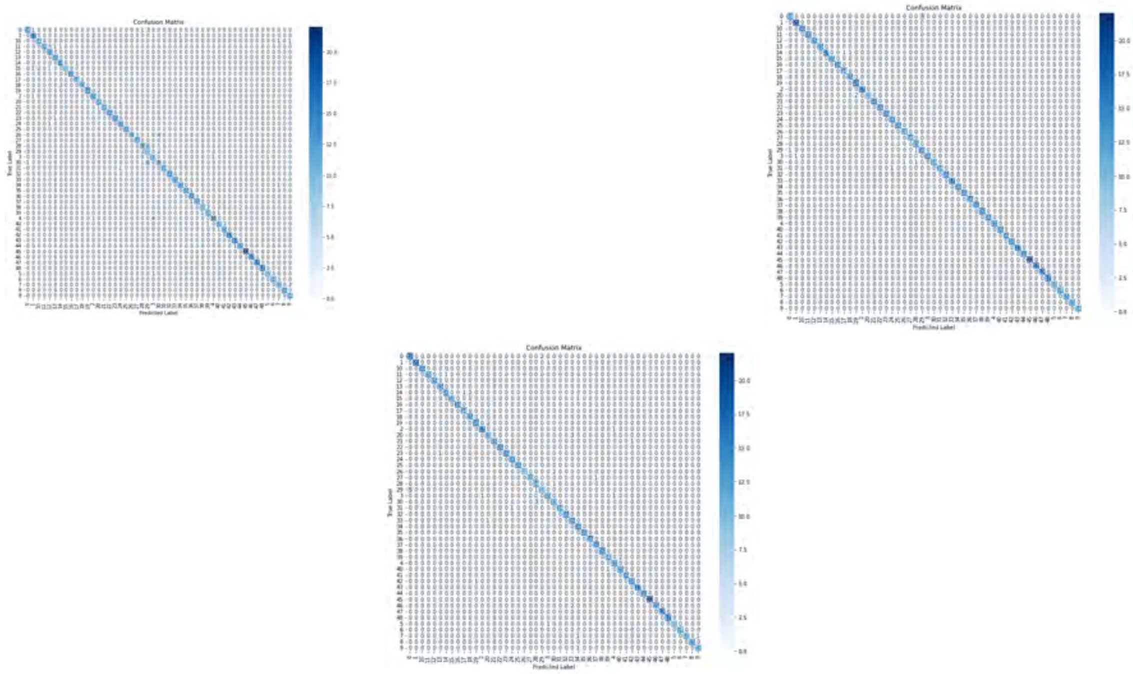


Figure 5.9: Confusion matrices for DenseNet, Inception, and ResNet on transfer learning

the other two models. Even so, off diagonal erroneous categorizations are reduced to a minimum and when appearing they tend to be densely distributed along the diagonal indicating less extreme errors in confusion among very different classes.

Roc curve:

The ROC curve of Resnet50 through transfer learning shows AUC usually is 1.00 for the majority of classes with class in particular having an AUC around. However, we will notice that many others are around which means that the ResNet50 model can classify most of the classes almost perfectly, even if class 1 is classified slightly less. Overall it has a great class discrimination with very few false positives. The overall performance of DenseNet121 is 1.0 in all classes indicating perfect classification with respect to other two models across all the classes. For this thermal dataset, after inducing transfer learning dataset the DenseNet121 model demonstrates a near perfect performance as every single class is correctly classified without any errors. This is shown with the AUC scores of all classes getting 1.00 which is a maximum value. For InceptionV3 in transfer learning we can see the classes 1-100 have an AUC of 1.00 like in ResNet50 and only class one has an AUC distinctly different which is 0.99. By analyzing the ROC curves we can say that the performance is very close to ResNet50's for class 1, we see a small drop. However, all other classes are stacked neatly, displaying an excellent general outstanding performance.

Details performance matrix: Based on the above comparison of the classification reports for the ResNet, InceptionV3, and DenseNet models, the transfer learning results in distinct performance differences in the machine learning classifications. While the latter three sections may be provided, this paper will focus on the differences in precision, recall, f1-score, and overall accuracy. InceptionV3 presented the best model and an overall accuracy of 97%. A macro average for precision, recall, and f1-score of 98% were obtained. Based on the AUC score of 99.8%, it is evident

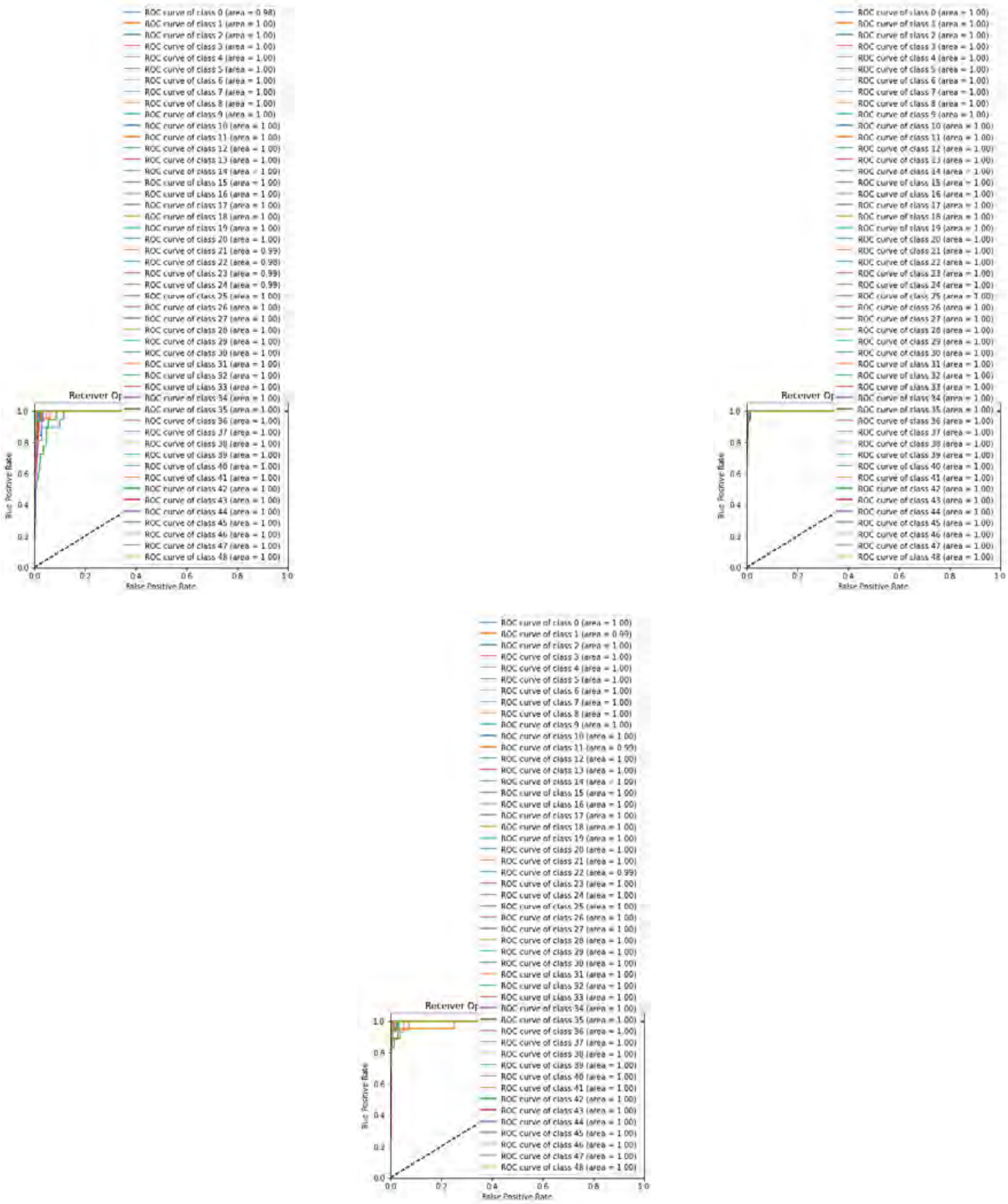


Figure 5.10: ROC curve of DenseNet, Inception, and ResNet on transfer learning

that the model was excellent at balancing out the class imbalances, suggesting that it would be the most reliable classification model for this task. The performance was

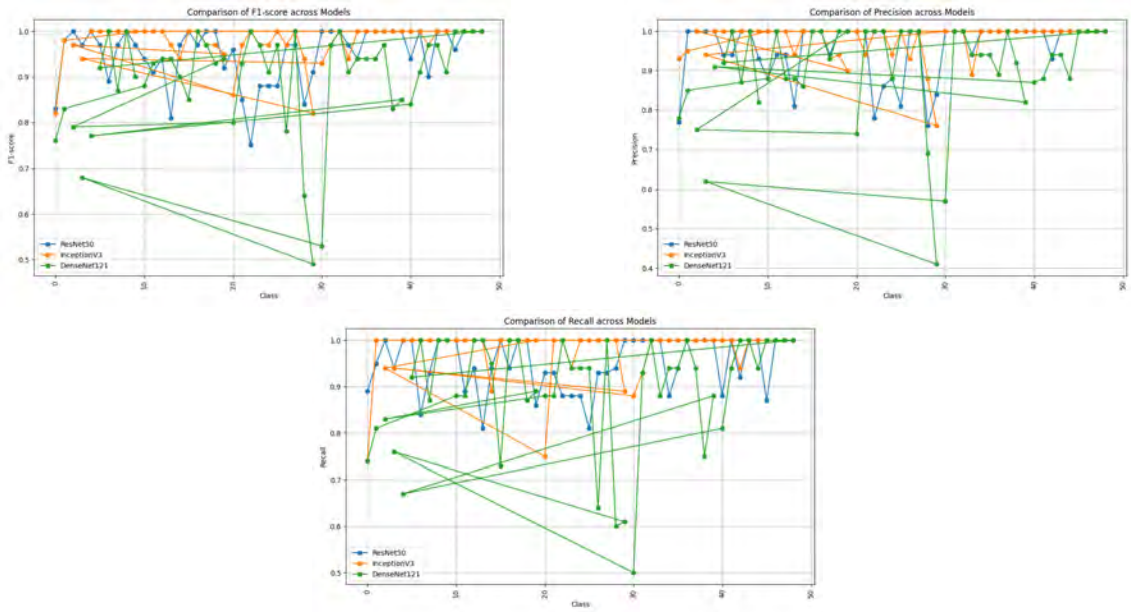


Figure 5.11: Comparison of F1-score, Precision and Recall across models of Transfer learning

outstanding for the different classes, and most of them attained perfect 1.0 for their precision, recalls, and f1-scores. A slight drop is noticed for class 0 with a precision of 0.93 and class 29 with a precision of 0.86. However, the act is consistent and very reliable. ResNet showed good results, and the model had an overall performance of 95%. Macros for precision, recall, and f1-score of 95% were obtained. The result was solid for nearly all models, and the AUC was 0.9990, indicating that it was excellent in maintaining a balance between accuracy and recall, making it an excellent transfer learning model.

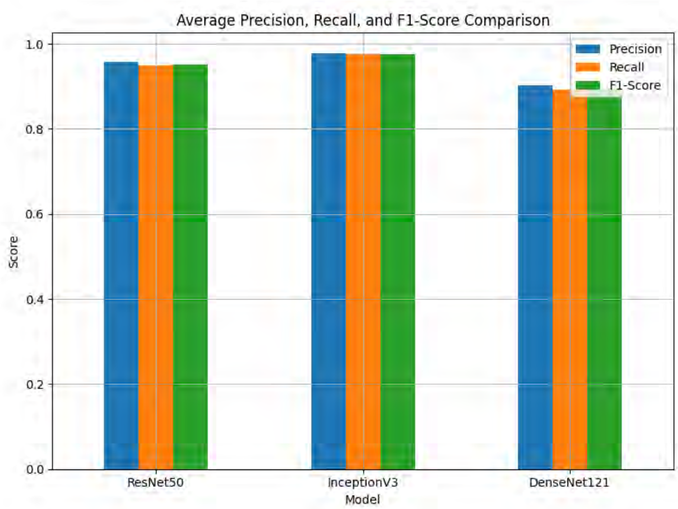


Figure 5.12: Average Precision, Recall and F1- Score Comparison of Transfer Learning

The good results are largely consistent with the model, and there are small variations in the precision, recall, and f1-score which is largely due to a small drop in its recall presented in classes such as 0 and 29.For DenseNet, the model showed the least performance, having an overall classification of 89% and the precision, recall, and f1-score of 89%. The results are variable in terms of precisions. The AUC of 0.9980 shows that it is a reliable transfer learning algorithm, but this was consistently less stable than it was for InceptionV3. Therefore, InceptionV3 was the most balanced and reliable classification model, followed closely by ResNet, and DenseNet was less likable due to the variations in classifications.

Table 5.1: Accuracy Analysis

Model	**Colormap Dataset Accuracy**	**Thermal Dataset Accuracy**	**Transfer Learning Accuracy**
ResNet50	98%	90%	95%
Inception V3	84%	92%	98%
DenseNet 121	81%	81%	90%

Chapter 6

Conclusion

This research was conducted for the betterment of people with a hearing impairment, it is our pivotal duty to communicate easily and should be accessible. Given two distinct datasets, we compared three state-of-the-art neural network models: ResNet, DenseNet and Inception. By performing extensive experiments and comparing appropriate features, we were able to uncover the advantages and limitations of each model in interpreting sign language gestures. Additionally, we used transfer learning techniques where possible thereby improving the model's accuracy to as high as 95% for ResNet; up to 90% for DenseNet and an exceptional figure of up to a massive 98 % in case of Inception. Together, these advances point toward the promise of deep learning for addressing long-standing challenges in hearing-impaired populations and provide a strong platform on which to build future research. We sincerely hope our work motivates more studies and creational endeavors aimed at producing efficient communication tools for the impaired in universal society, thereby promoting an environment of acknowledgement and understanding.

Bibliography

[1] C.-L. Huang and W.-Y. Huang, "Sign language recognition using model-based tracking and a 3d hopfield neural network," *Machine vision and applications*, vol. 10, pp. 292–307, 1998.

[2] T. Asada, Y. Yoshitomi, and R. Hayashi, "A human-machine cooperative system for generating sign language animation using thermal image," *Artificial Life and Robotics*, vol. 13, pp. 36–40, 2008.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, *Rethinking the inception architecture for computer vision*, 2015. arXiv: 1512.00567 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1512.00567.

[4] M. HeK, Q. RenS, *et al.*, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[5] L. Zheng, B. Liang, and A. Jiang, "Recent advances of deep learning for sign language recognition," in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, 2017, pp. 1–7.

[6] A. Das, S. Gawde, K. Suratwala, and D. Kalbande, "Sign language recognition using deep learning on custom processed static gesture images," in *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, IEEE, 2018, pp. 1–6.

[7] M. Hossen, A. Govindaiah, S. Sultana, and A. Bhuiyan, "Bengali sign language recognition using deep convolutional neural network," in *2018 joint 7th international conference on informatics, electronics & vision (iciev) and 2018 2nd international conference on imaging, vision & pattern recognition (icIVPR)*, IEEE, 2018, pp. 369–373.

[8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, 2018. arXiv: 1608.06993 `[cs.CV]`. [Online]. Available: https://arxiv.org/abs/1608.06993.

[9] G. A. Rao, K. Syamala, P. Kishore, and A. Sastry, "Deep convolutional neural networks for sign language recognition," in *2018 conference on signal processing and communication engineering systems (SPACES)*, IEEE, 2018, pp. 194–197.

[10] L. Y. Bin, G. Y. Huann, and L. K. Yun, "Study of convolutional neural network in recognizing static american sign language," in *2019 IEEE international conference on signal and image processing applications (ICSIPA)*, IEEE, 2019, pp. 41–45.

[11] M. S. Islam, S. S. S. Mousumi, A. S. A. Rabby, and S. A. Hossain, "A simple and mighty arrowhead detection technique of bangla sign language characters with cnn," in *Recent Trends in Image Processing and Pattern Recognition: Second International Conference, RTIP2R 2018, Solapur, India, December 21–22, 2018, Revised Selected Papers, Part I 2*, Springer, 2019, pp. 429–437.

[12] P. Maharjan, T. Bhatta, and J. Y. Park, "Thermal imprinted self-powered triboelectric flexible sensor for sign language translation," in *2019 20th International Conference on Solid-State Sensors, Actuators and Microsystems & Eurosensors XXXIII (TRANSDUCERS & EUROSENSORS XXXIII)*, IEEE, 2019, pp. 385–388.

[13] P. P. Urmee, M. A. Al Mashud, J. Akter, A. S. M. M. Jameel, and S. Islam, "Real-time bangla sign language detection using xception model with augmented dataset," in *2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, IEEE, 2019, pp. 1–5.

[14] D. Aich, A. Al Zubair, K. Z. Hasan, A. D. Nath, and Z. Hasan, "A deep learning approach for recognizing bengali character sign langauage," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2020, pp. 1–5.

[15] S. Hossain, D. Sarma, T. Mittra, M. N. Alam, I. Saha, and F. T. Johora, "Bengali hand sign gestures recognition using convolutional neural network," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, IEEE, 2020, pp. 636–641.

[16] M. J. Hossein and M. S. Ejaz, "Recognition of bengali sign language using novel deep convolutional neural network," in *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, IEEE, 2020, pp. 1–5.

[17] M. Oudah, A. Al-Naji, and J. Chahl, "Hand gesture recognition based on computer vision: A review of techniques," *journal of Imaging*, vol. 6, no. 8, p. 73, 2020.

[18] S. A. Shurid, K. H. Amin, M. S. Mirbahar, *et al.*, "Bangla sign language recognition and sentence building using deep learning," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, IEEE, 2020, pp. 1–9.

[19] D. Tasmere and B. Ahmed, "Hand gesture recognition for bangla sign language using deep convolution neural network," in *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, IEEE, 2020, pp. 1–5.

[20] M. S. Alam, M. Tanvir, D. K. Saha, and S. K. Das, "Two dimensional convolutional neural network approach for real-time bangla sign language characters recognition and translation," *SN Computer Science*, vol. 2, pp. 1–13, 2021.

[21] D. S. Breland, S. B. Skriubakken, A. Dayal, A. Jha, P. K. Yalavarthy, and L. R. Cenkeramaddi, "Deep learning-based sign language digits recognition from thermal images with edge computing system," *IEEE Sensors Journal*, vol. 21, no. 9, pp. 10 445–10 453, 2021.

[22] D. S. Breland, A. Dayal, A. Jha, P. K. Yalavarthy, O. J. Pandey, and L. R. Cenkeramaddi, "Robust hand gestures recognition using a deep cnn and thermal images," *IEEE Sensors Journal*, vol. 21, no. 23, pp. 26 602–26 614, 2021.

[23] S. Dhulipala, F. F. Adedoyin, and A. Bruno, "Sign and human action detection using deep learning," *Journal of Imaging*, vol. 8, no. 7, p. 192, 2022.

[24] M. J. Hasan, S. K. Nahid Hasan, and K. S. Alam, "Deep convolutional neural network-based bangla sign language detection on a novel dataset," in *Machine Intelligence and Data Science Applications*, V. Skala, T. P. Singh, T. Choudhury, R. Tomar, and M. Abul Bashar, Eds., Singapore: Springer Nature Singapore, 2022, pp. 157–168, ISBN: 978-981-19-2347-0.

[25] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A.-B. Gil-González, and J. M. Corchado, "Deepsign: Sign language detection and recognition using deep learning," *Electronics*, vol. 11, no. 11, p. 1780, 2022.

[26] K. K. Podder, M. E. Chowdhury, A. M. Tahir, *et al.*, "Bangla sign language (bdsl) alphabets and numerals classification using a deep learning model," *Sensors*, vol. 22, no. 2, p. 574, 2022.

[27] M. M. Damaneh, F. Mohanna, and P. Jafari, "Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using orb descriptor and gabor filter," *Expert Systems with Applications*, vol. 211, p. 118 559, 2023.

[28] A. Dayal, M. Aishwarya, S. Abhilash, C. K. Mohan, A. Kumar, and L. R. Cenkeramaddi, "Adversarial unsupervised domain adaptation for hand gesture recognition using thermal images," *IEEE Sensors Journal*, vol. 23, no. 4, pp. 3493–3504, 2023.

[29] D. Parashar, S. Thakur, K. B. Raju, G. B. Madhavi, and K. Sharma, "A deep learning-based approach for hand sign recognition using cnn architecture.," *Revue d'Intelligence Artificielle*, vol. 37, no. 4, 2023.

[30] S. Birkeland, L. J. Fjeldvik, N. Noori, S. R. Yeduri, and L. R. Cenkeramaddi, "Thermal video-based hand gestures recognition using lightweight cnn," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2024.

[31] S. S. Nathala, R. R. Yakkati, D. S. Breland, *et al.*, "A deep cnn-based hand gestures recognition using high-resolution thermal imaging," in *2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)*, IEEE, 2024, pp. 1–6.

[32] M. A. Rahaman, M. H. Ali, and M. Hasanuzzaman, "Real-time computer vision-based gestures recognition system for bangla sign language using multiple linguistic features analysis," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 22 261–22 294, 2024.

[33] M. A. Rahaman, K. U. Oyshe, P. K. Chowdhury, T. Debnath, A. Rahman, and M. S. I. Khan, "Computer vision-based six layered convneural network to recognize sign language for both numeral and alphabet signs," *Biomimetic Intelligence and Robotics*, vol. 4, no. 1, p. 100 141, 2024.

[34] M. J. Raihan, M. I. Labib, A. A. J. Jim, J. J. Tiang, U. Biswas, and A.-A. Nahid, "Bengali-sign: A machine learning-based bengali sign language interpretation for deaf and non-verbal people," *Sensors*, vol. 24, no. 16, p. 5351, 2024.

[35] S. Ahmed, M. Islam, J. Hassan, *et al.*, "Hand sign to bangla speech: A deep learning in vision based system for recognizing hand sign digits and generating bangla speech. arxiv 2019," *arXiv preprint arXiv:1901.05613*,