Human Activity Recognition - A comparative study using Traditional and NextGen Transfer Learning Pre-trained Models

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
SOFTWARE ENGINEERING

Submitted by

JITEN SUTAR (2K22/SWE/08)

Under the supervision of

MS. SHWETA MEENA

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

MAY, 2024

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

We wish to express our sincerest gratitude to Ms Shweta Meena for his continuous guidance and mentorship that he provided us during the project. He showed us the path to achieve our targets by explaining all the tasks to be done and explained to us the importance of this project as well as its industrial relevance. He was always ready to help us and clear our doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

Place: Delhi — Jiten Sutar

Date: 24.05.2024 — (2K22/SWE/08)

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Jiten Sutar, Roll No's –2K22/SWE/08 students of M.Tech (Software Engineering), hereby certify that the work which is being presented in the thesis entitled "Human Activity Recognition - A comparative study using Traditional and NextGen Transfer Learning Pre-trained Models" in partial fulfilment of the requirements for the award of degree of Master of Technology, submitted in the Department of Software Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from Jan 2024 to May 2024 under the supervision of Ms. Shweta Meena.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor (s)

DEPARTMENT OF MECHANICAL ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "Human Activity Recognition - A comparative study using Traditional and NextGen Transfer Learning Pre-trained Models" which is submitted by Jiten Sutar, Roll No – 2K22/SWE/08, Department of Software Engineering, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Ms. Shweta Meena
Assistant Professor

Date: 24.05,2024

Department of Software Engineering, DTU

Abstract

Human Activity Recognition (HAR) is very vital in appreciating human demeanor and finds applications in healthcare, sports analytics, and surveillance systems. Increasingly the data driven insights are being utilized HAR plays a key role in identifying patterns, trends and anomalies associated with human activities. The use of machine learning and deep learning techniques has helped to improve significantly HAR methodologies leading to higher accuracy and efficiency. This study provides an extensive insight into traditional and advanced transfer learning pre-trained models for exploring intricacies of HAR.

Each of the different model architectures in the research was assessed in depth by this evaluation, with its own strengths and capabilities. VGG16 or VGG19 and EfficientNetV2S, Xception are examples of old pre-trained models which would be compared with ConvNeXt frameworks such as ConvNeXtSmall, ConvNeXtBase, ConvNeXtLarge, and ConvNeXtXLarge. The main objective of this study is to comparatively analyze the efficiency of these models using human activity recognition. The benchmarking used a well-curated dataset that involved 12000 images annotated and classified into fifteen activities The Kaggle dataset sourced is useful for evaluating any performance changes made to different pretrained models. In order to avoid partiality and control external factors; like biases each model had exactly the same number of layers as others.

The experiments are carried out in the Google Colab environment, which is cloud-based and therefore allows for extensive experimentation and analysis.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Human Activity Recognition (HAR) is one such emerging field of Artificial Intelligence dealing with the automatic detection and classification of the tasks taken care of by men. This kind of field is important because of the applications of the same in various sectors. HAR can track health-wise movement of patients to ensure safety and wellness, detect falls, and for rehabilitation programs. This means that fitness tracking applications, when used in combination with real-time feedback, can provide performance improvements. Smart home systems put together the components of HAR to create better experiences for users. They automate homes in response to the residents and their activities[8]. These may involve dimming or brightening lights or controlling the climate whenever someone comes into or leaves a room. Surveillance systems, too, leverage the capability of HAR to auto-analyze footage, be it for noticing and responding to deviant behavior or suspicious activity.

The introduction of machine learning and deep learning-based algorithms to activity recognition techniques has brought this renaissance in HAR, where much better accurate and reliable models have been developed. Previously, the backbone of HAR was built based on traditional machine-learning-based models, particularly Support Vector Machines (SVMs) and Random Forests, using handcrafted features that were extracted from the sensor data to discriminate between various activities. These algorithms have worked really well with a solid mathematical foundation and applicability for a very wide range of classification tasks[12]. For instance, SVMs find an optimal hyperplane that passes to separate data points from various classes with the maximum margin. They are powerful tools for binary classification. Random Forests, on the other hand, are ensemble methods to build many decision trees and take the mode of classes as output for classification tasks. They certainly advance the problem of overfitting and low accuracy because several models are combined.

However, with these traditional approaches, a lot of domain expertise is always required to select and engineer the desired features from raw sensor data. It is done in a very laborious and crucial way because the quality and appropriateness of the features determine how well the model performs. Moreover, traditional algorithms may also suffer from the complexity and high dimensionality in modern

sensor datasets. Such datasets may likely contain high-dimensional sample points, collected from several sensors, that follow closely the nuances of user activity, a process that classical methods find hard to disentangle properly for suitable classification[9].

Limitation of conventional machine learning approaches brought acceptance of deep learning strategies in HAR. Remarkable properties of deep learning models, particularly CNNs and RNNs, are for automatic feature learning from raw data. Since the models are highly dimensional by themselves, they will have the capacity to capture temporal dependencies and spatial hierarchies in sensor data, hence increasing classification performance.

CNNs are designed to solve spatial patterns and are well adapted for local-correlation-type data processing, such as tasks involving images or sensor signals represented in time-series data. RNNs and their more advanced versions, such as LSTM, perform the task of capturing temporal dependencies, especially those sequence data where activities flow from one to another.

Although traditional machine learning algorithms, such as SVMs and Random Forests, have been an essential source of innovation in the very early development of HAR, this trend is rapidly altering with recent moves toward deep learning. This becomes really very critical because deep learning models can automatically represent most of the relevant features into a dataset characterized by complex, high-dimensional distributions, thus reducing the necessity for expert domain knowledge, making these systems more accurate and reliable. This shift is driving capability and applicability in HAR, opening the floodgates of innovation for healthcare, fitness, smart homes, and surveillance systems.

Deep learning techniques—especially Convolutional Neural Networks (CNNs)have propelled human activity recognition into higher pedestals. Most enabling features in CNNs are their automaticity for feature extraction and hierarchical features from raw data without any manual engineering. This has moved CNNs from an obscure field almost two decades ago to positions of prominence within HAR; they continuously produce state-of-the-art results and surpass most benchmarks of traditional machine learning algorithms.

There are some unique merits in using the deep learning methodologies in the field of HAR. First, the introduced deep learning models alleviate qualitative difficulties to a great extent when feature engineering manually. Feature engineering, as in any manual method, requires great effort and deep domain knowledge for identifying and extracting meaningful features from sensor data. This process is labor-intensive and moreover prone to human errors and biases. CNNs make it possible to learn the patterns in complex data automatically and thus enable development much more efficiently and effectively.

Deep learning models, on the other hand, have been consistently showing much higher accuracy in recognizing activity compared to previous traditional methods. These CNN models are deeply structured to perfectly learn and differentiate a wide range of patterns in input data, including very subtle ones or those hard to perceive. This inherently gives better classification among these activity classes, as the model can sense variations and nuances that escape traditional algorithms.

Thirdly, deep learning frameworks have the ability to be trained on unlabeled data. In such scenarios as HAR applications, it is impossible or generally, in most cases, too expensive to obtain fully labeled datasets. Semi-supervised and unsupervised learning techniques allow models to learn from vast pools of unlabeled data to improve their learning process and performance. In this way, the ability to learn from a small set of labeled data and big pools of unlabeled data makes deep learning models very flexible and scalable.

Furthermore, such deep learning models generally maintain efficiency across diverse datasets and incorporate variations that arise from individual differences, device models, and environmental conditions. Traditional machine-learning algorithms find it difficult to generalize across conditions because they strongly depend on the features engineered for a specific dataset. CNNs and other deep-learning models can show better generalization behaviors because of the learned hierarchical feature representations. Thus, this ensures that the models are consistently performed well without regard to variance in the input data.

In addition, since deep models can be fine-tuned toward any kind of application within the context of HAR, they are therefore very flexible in use. Transfer learning, on the same note, can help in leveraging a pre-trained model on a large-scale dataset and adapt it to a new but related task. This enables reuse of models trained with huge datasets for tasks involved in HAR, yet at the same time reducing demand on computational resources and training time largely.

The use of deep learning methodologies, especially CNNs, has really revolutionized the field of HAR, where automated feature extraction could now be done without relying on manual feature engineering. The benefits of deep learning methodologies for HAR are diverse: they simplify the process of development, improve precision in recognizing human activities, learn from unlabored data, and ensure generalization robustness over various datasets. Together, these advantages place deep learning at the forefront of research and applications in HAR, paving the way for more advanced, robust, and scalable activity recognition systems[15].

A major shortcoming for deep learning approaches, and more specifically for Convolutional Neural Networks, is their demand for massive amounts of training data. The weights must be tuned using large training datasets in order to improve their capability of generalization in the future. However, producing massive labeled datasets remains a challenging task in many fields. These problems are solved by Transfer Learning (TL) at this point.

1.1 Problem Statement of Dissertation

Human Activity Recognition (HARR) is an important application domain in artificial intelligence and computer vision tasked with the automatic recognition and classification of what humans do. It is paid prime importance in health, fitness, smart home systems, and surveillance domains. Current methods for recognizing human activity depend on pre-trained models with manual feature extraction; however, the existing approaches from both challenges resulted in accuracy and

efficiency when dealing with complex datasets. Each of these architectures, at some level, includes the use of ConvNeXt. Combined with the drive for more accurate and effective performance in HAR, these architectures advance over classical pre-trained models. This dissertation shall seek to do full comparison between traditional pre-defined transfer learning models like VGG16, VGG19, EfficientNetV2S, and Xception over the newer architectures like ConvNeXt. Our goal is to assess their performance through a 12,000-image dataset, compiled from fifteen different classes of activities, and express new better capabilities of the ConvNeXt architectures related to HAR.

1.2 Overview of the research objectives of the Dissertation

The issue concerning this dissertation is that traditional pre-trained models are not that effective in recognition considering the human activities. The traditional models, including VGG16, VGG19, EfficientNetV2S, and Xception, have overlapped many applications, including HAR, in almost any task of image processing. However, such models often fail to identify human activities because of their complexity and variability in datasets or the approaches taken to get those datasets. They are further challenged by the correct extraction and interpretation of the subtle patterns and features inherent to human activity that are necessary for reliable recognition.

ConvNeXt architectures are realized with newly designed convolutional layers that have been purposely designed to capture patterns while being intricate and thus promising. This dissertation therefore leverages a dataset of 12,000 images spread over 15 different classes of activities into considering an extensive evaluation of how these newer models perform compared with traditional models. From the results, it is evidently clear that ConvNeXt models, more so the ConvNeXtLarge, have done a remarkable improvement in the accuracy and robustness involved in HAR tasks.

1.3 Transfer Learning in Machine Learning Research

Transfer learning has found wide importance of late in the current days in machine learning research. It applies already learned knowledge from source domain either for improving the efficiency of learning or, it is especially known not to need a large number of labeled examples for target domain. Unlike traditional learning frameworks, transfer learning does not require that the data for training and testing come from the same domain. Such flexibility allows models to apply the knowledge gained by well-analyzed large datasets from one domain to different but related tasks, hence significantly enhancing their performance in cases of scarcity of labeled data within the new domain.

Consequently, transfer learning minimizes the need for building target domain models from scratch and reduces data scarcity problems during training. Not only would this reduce the resources and time needed to train the computational model, but it would also offer a transferred ability of generalization over new tasks in the model by using prior knowledge.

From the CNN transfer learning literature, three dominant categories can be identified: fixed feature extraction, fine-tuning with frozen layers, and pre-trained models. Each category comes with its own pros and cons, and is uniquely adapted to the specifics of each target domain.

1.3.1 Fixed Feature Extraction:

The pre-trained CNN acts as a fixed extractor of features. Features are extracted from input data using the convolutional layers of the pre-trained network, which are then fed to a new classifier. This is beneficial since it would leverage the salient feature extraction ability of CNNs without much loss in retraining. However, this might work less effectively if features learned from the pre-trained model are not that relevant to the new task.

1.3.2 Fine-tuning and Layers Freezing:

In this method, the pre-trained network is partially re-trained on the new dataset. In such a transfer learning scenario, few samples are available; therefore, some layers are frozen from the beginning of the network, and the remaining part is fine-tuned to fit the new task. Such a methodology balances between pre-trained features and adaptation to data-specific characteristics. It is computationally more expensive compared to fixed feature extraction but can result in better performances compared to cases in which the source and target domains are slightly related[18].

1.3.3 Using Pre-Trained Models:

This category includes using directly models pre-trained on large and generic datasets, such as ImageNet, as a basis to approach new tasks. Pre-trained models such as Xception, DenseNet, and VGG16 are the base pillars of transfer learning in HAR research. The models are training on very large datasets, so the features learned by them are very diverse and can then be transferred to many kinds of tasks. The main advantage of such models is that they come with a rich set of learned features, potentially fine-tuned further for new applications. The bad part is that most might need a major adaptation to perform even close to optimally in tasks quite different from their original training domain. Transfer learning provides a pragmatic solution for challenges that arise due to the need for large, labeled datasets, and more so in deep learning that is energy based. By enabling models to leverage knowledge that is already available, transfer learning accelerates the training process and boosts the model's performance in low-data regimes. This makes them valuable tools in developing HAR systems where

collecting large labeled datasets can be particularly difficult. The potential of transfer learning in improving HAR research and applications is demonstrated by the application of sophisticated, pre-trained models such as Xception, DenseNet, and VGG16, among others, in very robust and adaptable frameworks to myriads of activity recognition tasks.

1.4 Overview of the Study

This paper explores several models in the domain of human activity recognition, such as VGG-16, Xception, and ConvNeXt, for measures and generally for accuracy. As always, critically probing how they fair, these models were further fine-tuned in the Keras library for adopting experiments and analyses within the Google Colab workspace using well-established mechanisms of transfer learning without undergoing any challenge form. It gives researchers ways to effectively use powerful pre-trained models, thus allowing for significant improvement in experimentation and evaluation[6].

The article is arranged in such a manner that one ensures a look has been taken at all key aspects discussed. To do that, first, the depth of the methodology used within the process of experimentation is looked into. This entails how the models were set up, trained, and the datasets used on the experiments, together with which specific transfer learning methods were applied[19]. Basically, this let the researchers take advantage of some pre-acquisition knowledge that the already pre-trained models had, thus avoiding the hassle needed for massive labeled data and speeding up the entire process.

All of these were very carefully designed with the utmost concern to ensure that each model's performance is measured correctly. This has been achieved by comparing the performance of different models in showing/recognition of different human activities. Here, the measurement of how well the model can classify activities correctly using sensing data is emphasized in a really big way, since accuracy is such an important performance metric in HAR[16]. Application of the Keras library through the Google Colab platform was flexible and powerful for implementing and testing these models, thereby allowing for quick model runs and experiments without much computational constraint.

The paper then lays out the detailed methodology before proceeding to present the obtained findings from the study. These range from quantitative results in terms of accuracy scores of different models to qualitative insights on how well each model is conditioned. The analysis of these results gives important information about the current state of HAR research: strengths and weaknesses of different deep learning approaches.

This paper will also consider possible implications for today's HAR research. The fact that deep learning models dramatically change the landscape of human activity recognition when boosted by transfer learning techniques underlies this paper. It has been said that transfer learning became a very important weapon in this domain to achieve high performance even with low quantities of annotated data. This latter aspect is important because for many HAR problems, collecting

large-scale datasets, accurately labeled, is both an expensive and laborious task[7].

Use of pre-trained models, for example, VGG-16, Xception, and ConvNeXt, show how powerful using existing well-trained architectures could be in increasing efficiency and accuracy of HAR systems. These models have been trained on very large representations, including the ImageNet dataset, hence providing a strong baseline for transfer learning techniques, which will support the researcher in developing a robust HAR model fast and with less data[11].

Thus the paper delves further into the various models in the HAR domain, with more emphasis on the accuracy and performance metrics of the models. This experimentation was efficiently and effectively conducted by the researchers through the use of Keras pre-defined transfer learning mechanism realized in the Keras library onto Google Colab[1]. These results proved the enormous potential of deep learning and transfer learning in transforming HAR research, adding another mile on the road toward more accurate and robust activity recognition systems. This work not only pushes the state of the art in the understanding of HAR but also sets a very solid step in the area, proving in practice how deep learning can truly revolutionize the way human activities are recognized and analyzed.

Chapter 2

RELATED WORK

In this chapter, all the related works related to human activity recognition (HAR), traditional transfer learning pre-trained models and new pre-trained models are elaborated.

2.1 HAR Based on Traditional Machine Learning and Deep Learning

Human activity recognition is a form of classification in which human activities are detected and classified according to the data captured in video recordings, data from sensors, and images. There are many approaches to HAR most are traditional methods under which handcrafted features are stacked together using sensor data both in the time and frequency domains. Such features include standard deviation, Pearson coefficient, harmonic mean, among others, which then form input to several recognition models.

Some of the traditional machine learning algorithms that have been discussed and applied to HAR include random forests, decision trees, support vector machines, and K-nearest neighbors. Thus, these models classify between the several activities using the handcrafted features in providing the characterization of the features concerning the activities. For example, random forests and decision trees create a chain of decision rules based on the features, whereas SVMs find and determine an optimal hyperplane that separates the classes of activities. KNN, in contrast, classifies activities subject to the nearest labeled examples in the feature space.

Great results have been achieved with traditional methodologies, and there are a large number of studies supporting this view. Huge precisions in recognizing a vast range of human activities were reached using traditional methods tuned to find the most relevant aspects of the sensor data through feature engineering. However, either domain expertise or subtleties within the data can easily be overlooked, which is a potential drawback for this approach. However, these methods again were able to establish a solid platform for further research in HAR.

Traditional HAR approaches effectively work with handcrafted features from sensor data; that is, they leverage algorithms such as random forests, decision trees, SVMs, and KNN in order to obtain reliable recognition of activities. These successes reveal the value of these methods even as the field increasingly leans toward more automated feature extraction and advanced deep learning techniques.

For instance, several studies have proven the effectiveness of traditional machine learning approaches in HAR by using hand-crafted features and employing various classifiers. For example, Casale et al. (2011) [4] achieved reasonable accuracy in recognizing basic daily activities by using a computationally efficient feature model through a random forest approach. The current work has established some key attributes of random forests: handling large datasets in a feature-rich setting in a computationally efficient way.

In a study by Casale et al., random forest was used for encoding the recognition of basic daily activities. In doing so, they took full account of computationally efficient features in their work to balance the fine line between accuracy and processing time. In this work, the use of the random forest algorithm, known for its robustness in handling large datasets, fits the task perfectly. They used basic features, which ensured that the model they built could be implemented in real-world applications without the use of sophisticated computational resources. The study showed that even with primary features, random forests can give acceptable performance in terms of accuracy for discriminating activities such as walking, sitting, standing, and running. This work heralds that potential to be used in practical HAR applications where efficiency is crucial.[4].

In the experiments conducted by Ayman et al., the PAMAP2 dataset was used, which includes a large number of physical activities recorded with different sensors. They used feature selection techniques to fuse sensor data for enhanced activity recognition. The feature selection process should identify the most relevant features with a simultaneous decrease in dimensionality and improvement in performance. They made full use of the complementary information in data from several types of sensors, such as accelerometers, heart rate monitors, and gyroscopes, to enhance classification accuracy. The random forest classifier was used for this application, assuming that a classifier could be applied to diverse and high-dimensional data. Their approach has shown that combining feature selection with sensor fusion can improve the accuracy of HAR models significantly[3].

Mekruksavanich et al. proposed a very complete framework for activity recognition using accelerometer, gyroscope, and surface electromyography data. These methods contribute together by using a decision tree model to interpret and combine the data from more than one source of sensor input. The simplicity of the decision tree algorithm and its high interpretability make it one viable choice for this multisensor setup. They would further be able to record a large amount of activity-related information, such as the dynamics of movements and muscle activity, through data from varied types of sensors. In this case, their framework was able to classify different exercises and physical activities quite competently; thus, decision trees were sufficiently handy and flexible to use in multi-sensor HAR applications[14].

Arif et al. based their work on extracting time-domain statistical features from accelerometer data for purposes of physical activity classification. Such features were such as mean, standards deviation, skewness, and kurtosis calculated from the accelerometer signals. Such features did disclose information regarding the characteristics of activity quite well: for instance, walking might show a regular periodic acceleration, whereas activities like sitting would show minimum variations. Then, these time-domain features were used to recognize activities predicted from the classification algorithms. Their study emanated in very optimistic results and showed that even simple statistical features, if properly extracted and used, could discriminate across classes of different physical activities. This highlighted the importance of properly engineered features in traditional methodologies of HAR[2].

In recent decades, work in the field of HAR has gradually moved to deep learning methodologies because they enable the automatic learning of distinguishing features directly from raw sensor data. And it is the case that remarkable improvements and achievements in the field of deep learning, focused on matters concerning object tracking, image classification, and speech recognition, have sufficed to bring about a shift toward deep feature-based methods. 1D and two-dimensional (2D) Convolutional Neural Networks (CNNs) have been means of growing importance in the field of HAR, yielding a top performance compared to those of traditional approaches that mostly rely on hand-crafted features.

CNNs are particularly useful for HAR because they are capable of efficiently processing raw sensor signals to extract hierarchical characteristics. In 1D CNNs, convolutional operations are applied to input data in the form of time-series data, which is a correct signal representation for typical sensor data used in HAR.

Thus, these models can capture local dependencies and temporal patterns that correspond to the suitability of recognizing complex activities. On the other hand, 2D CNNs are applied after transformation of sensor data into a 2D format, such as spectrograms or images, enabling the model to take advantage of spatial relationships within the sensor data.

The development of human activity recognition, recently increased by the use of CNNs, applies models that are more sophisticated, including ensemble methods with multiple CNN streams. These ensemble architectures are superior to their single-stream analogs in a way that they can combine different perspectives and features from multiple data representations. Such a multisource approach indeed increases both robustness and accuracy in the activity recognition models. More in general, recurrent neural networks are one of the most widely known and frequently implemented approaches to HAR. The Long Short-Term Memory subtypes are common because they work well with practically all sequential data. They are designed to capture long-range dependencies and temporal correlations in time series, which makes them applicable in tasks related to sequences. A case in point is HAR.

For a concrete example, HAR through smartphone data was performed with the help of a five-layered stacked LSTM network by Ullah et al., 2019. This deep architecture of LSTM can effectively capture the temporal dynamics relevant to

sensor signals and hence aid in performance improvement for the recognition of different activities. Still, it has the power to model complex dependencies right in time to distinguish between the activities underway[17].

Likewise, Hernandez et al. documented the benefits of using a BLSTM network for HAR. The key feature of a BLSTM is that it processes data in both forward and backward directions while at the same time being able to capture information from both past and future time steps[10]. In fact, such a feature can be very useful within HAR, likely to provide context from previous and subsequent data points towards further differentiating otherwise similar activities. For example, they found that a BLSTM was significantly better in distinguishing such highly related activity classes as walking upstairs and downstairs, which previously used to be identified as challenging distinctions to bring out because of their close resemblance in patterns.

All these advanced studies highlight once again that deep learning is the core technology changing HAR. Deep learning models can automatically learn features from raw data, and therefore can avoid the time-consuming and domain-knowledge-reliant process of engineering hand-crafted features. Therefore, the ability of CNNs and LSTMs to capture very complex patterns and dependencies in time has significantly enhanced the accuracy and robustness of HAR systems.

Future improvements on HAR come from integrating deep learning approaches with ensemble methods and advanced RNN architectures, including LSTMs and BLSTMs. These further pave the way for developing sophisticated, reliable recognition applications that could rightfully be used in application domains such as health and fitness, smart homes, and surveillance.

In this context, research efforts in deep learning for HAR have produced models able to learn from raw sensor data. Thus, various techniques comprising 1D and 2D CNNs, LSTMs, as well as BLSTMs are highly effective and consistently outperform methods based on traditional handcrafted features. The continuous evolution and development of these deep learning methods, in the context of huge benefits in HAR, put the studies forward and bring improvements for enhancing the precision and usefulness of an activity recognition system.

These are diversified methodologies and techniques that underline the diversified nature of HAR research. However, the use of all these models and techniques tends to come in the way of challenges associated with recognition and classification of human activity. More specifically, by developing models for deep learning techniques like convolutional neural networks and long short-term memory networks, one can learn complex features and temporal dependencies directly from raw sensor data.

In this respect, the growing evolution of activity recognition research, empowered by such advanced methodologies, heralds further novel advancements to the classification and understanding of various human activities. The more researchers continue to explore and refine such techniques, the better the HAR systems considered become in obtaining improved accuracy levels, robustness, and handling of divergent and complex datasets. This development is highly important for the development of ambient intelligence applications in various do-

mains such as healthcare, fitness, smart homes, and surveillance. The precision with which activities are recognized carries great potential for enhancing users' experiences and improving the overall system performance.

Finally, the heterogeneity in methodologies and techniques within the researchers working on HAR makes the field multidimensional with great potential for improvement. Concurrent models involving a number of CNN streams and where RNNs have been applied until now, especially LSTMs and BLSTMs, prove to be very effective toward the enhancement of the recognition of activities. These improvements will serve to illustrate an area that can only be actualized with more sophisticated machine-learning models so that, in return, more sophisticated and reliable HAR systems can be developed.

2.2 Transfer Learning Based Method for Human Activity Recognition

Two main approaches have been widely researched for the field of Human Activity Recognition:

- Hand-crafted feature-based methods
- Deep learning-based methods

Furthermore, some methods combine properties from both the techniques into a hybrid model to leverage the best of both worlds.

Feature-based handcrafted approaches The handcrafted feature-based methods correspond to creating an algorithm that manually extracts features from raw data used as input to other, more advanced machine learning tools. Those features may come from a time or frequency domain, including statistics such as mean, standard deviation, and skewness, or complex features like wavelet coefficients and spectral entropy. Some of the algorithms that are usually used in conjunction with handcrafted features are Support Vector Machines (SVMs), Random Forests, k-Nearest Neighbors (k-NN), and Decision Trees. While these methods require much domain expertise to identify the most informative features, they have had a good performance record in HAR tasks.

2.2.1 Deep Learning-based Methods

The most important development in HAR recently, after the invention of deep learning, is that of the CNN and RNN architectural formulations. It revolutionized this area of deep learning in HAR tasks since they can automatically extract features from raw data. Hierarchical feature representations are learned by such models, thus capturing complex patterns and dependencies without having to do explicit feature engineering.

Convolutional Neural Networks (CNNs)

It have been shown to be highly effective in the treatment of HAR because of their property pertaining to the ability to capture spatial hierarchies in data. Convolutions are applied to time-series data in 1D CNNs, whereby the category is useful for processing sensor signals. Most of the time, the sensor data can be transformed with some kind of reshaping technique into a 2D form, thereby enabling the use of 2D CNNs for exploiting spatial relationships. One of the key arguments for using CNNs lies in the fact that they are capable of capturing local patterns and are richly available within the sensor data.

Recurrent Neural Networks (RNNs)

: Long Short-Term Memory (LSTM) networks, a subclass of RNN models, have set state-of-the-art performance in modeling sequential data. LSTMs capture long-term dependencies and temporal dynamics, which are critical in recognizing activities unfolding in time. This has further been improved with a new architecture called bi-directional LSTM, which processes features in both forward and backward directions to furnish more accurate context that is temporal in nature.

2.2.2 Hybrid Techniques

Some of these HAR approaches thus constitute a hybrid of the hand-crafted, feature-based technique and deep learning. Hybrid models leverage this by using features that are generated manually as additional inputs in deep learning models or combining traditional machine learning algorithms with neural network layers to take advantage of the robustness and interpretability that handcrafted features may lend, while still benefiting from the strong representation learning of deep learning models.

2.2.3 Transfer learning in HAR

Transfer learning is one of the most effective ways to deal with HAR problems using pre-trained models, which had been initially trained on large image datasets, including ImageNet. Such models have already been pretrained so as to learn the rich feature descriptor from voluminous data and are fine-tuned or adapted in order to handle HAR tasks. The following are considered the principal advantages of transfer learning:

Lowers Training Time

This is so because a pre-trained model has already gone through the learning process of useful features, which will be transferred to the HAR domain.

Improved Small Data Performance

Transfer learning is most useful when the labeled data is very small, as pre-trained models on large datasets are very strong at feature extraction.

Enhanced Generalization

Pre-trained models typically generalize better for new tasks because they have undergone extensive, varied large-scale training data.

These breakthroughs clearly illustrate the current evolution within HAR methodologies, with deep learning and transfer learning emerging as potent tools in the development of more accurate activity recognition systems. On an equal footing, the next stage in this evolution is the proper understanding and classification of human activity, which will lead to a wide spectrum of practical implementations in the areas of health care, fitness, smart homes, among others.

In classification, both the spatial and temporal templates fine-tune a pre-trained VGG-16 model. Such a method uses transfer learning, where the weights of a model are actually initialized, first trained on huge image datasets like ImageNet, and then fine-tuned to work for the specific task of activity recognition for in-home residents. This method enhances the generalization and performance of models through avoidance of training deep learning architectures from scratch and appropriates high-dimensional feature representations.

Other descriptors like GHI and TAGBM are added to this system to make the model specific for human activities of subtle difference. After that, spatial and temporal templates regulate classification of the model, which has made the model clear to identify different forms of activities.

This has been proven effective in experiments on benchmark datasets like KTH and UCF Sport actions, which makes it efficient and robust for activity recognition tasks of totally different natures. By combining the ideas from domain adaptation and transfer learning with these novel descriptors, Zebhi et al. have successfully pushed this paper to lead the research direction toward HAR, opening new doors for human activity understanding. Herein lies the potential for this innovative approach to be inspiring towards further advancements in recognition methodologies of activity and resulting in far more sophisticated models enabling the solution of very real world problem challenges.

Title of the Paper	Year	Author(s)	Findings
A multi-class classification approach for Human Activity Recognition based on accelerometer data	2011	Casale et al.	Achieved reasonable accuracy in recognizing basic daily activities using a computationally efficient feature model and a random forest approach. Demonstrated the effectiveness of random forests in handling large, feature-rich datasets.
Efficient activity recognition from low-level sensor data	2019	Ayman et al.	Used feature selection techniques to fuse sensor data for enhanced activity recognition, improving accuracy significantly by combining feature selection with sensor fusion using the PAMAP2 dataset.
An exercise recognition framework using multi-sensor data	2020	Mekruksavanich et al.	Proposed a framework for activity recognition using accelerometer, gyroscope, and surface electromyography data with a decision tree model, effectively classifying different exercises and physical activities.
Physical activity classification using time-domain statistical features	2015	Arif et al.	Extracted time-domain statistical features from accelerometer data for activity classification, showing that simple statistical features can effectively discriminate different physical activities.
Human Activity Recognition using stacked LSTM networks	2019	Ullah et al.	Demonstrated the effectiveness of a five-layered stacked LSTM network in capturing temporal dynamics relevant to sensor signals, improving performance in recognizing different activities using smartphone data.
Benefits of BLSTM networks in Human Activity Recognition	2019	Hernandez et al.	Showed that BLSTM networks, processing data in both forward and backward directions, significantly improve the accuracy of distinguishing similar activities, such as walking upstairs and downstairs.
Various deep learning models and techniques for Human Activity Recognition	2010s-2020s	Multiple researchers	Highlighted the effectiveness of CNNs, LSTMs, BLSTMs, and ensemble methods in automatically learning features from raw sensor data, consistently outperforming traditional methods relying on handcrafted features.

Table 2.1: Summary of Related Work

Chapter 3

RESEARCH METHODOLOGY

Methodology In this section, we present a chapter explicating an in-depth analysis—including the technical underpinning—of our methodology for Human Activity Recognition. We hereby describe the technical backbone of our approach, including transfer learning and advanced deep learning architectures; particularly, the ConvNeXt model. An expository discourse on the methodologies used is presented starting with the deep technical background, through descriptions of the datasets used and the methods applied.

The base for our methodology in HAR lies in the concept of the usage of transfer learning and models under ConvNeXt. Transfer learning became the foundation that totally revolutionized machine learning when it allowed these models to apply big pools of pre-trained knowledge, hence greatly reducing the necessity to source huge labeled datasets for new tasks. This is especially useful in the context of HAR because the collection and labeling of such huge amounts of activity data can thus be expensive and time-consuming. Transfer learning allows us to fine-tune neural network configurations that have already been pre-trained on huge, large-scale datasets like ImageNet. Indeed, architectures such as VGG16, GoogleNet, and Residual Networks have learned in-depth the feature representations shared by millions of images that could be fine-tuned for our special-purpose HAR tasks, thus making it possible to increase performance and generalize to more complex cases.

ConvNeXt models are a new development in neural network architecture that can take parallel input data through convolutional channels. It smoothens and enhances feature extraction and representation learning since it captures both the local and global patterns of the data. The parallel pathways imposed on ConvNeXt models allow for comprehensive analysis of input data, which becomes very useful in the case of complex tasks like HAR.

We use two popular benchmark datasets, KTH dataset and UCF Sport Action dataset, with four scenarios each. The KTH dataset consists of video sequences from six types of human actions: walking, jogging, running, boxing, hand waving, and hand clapping, performed by 25 subjects under four different scenarios in terms of the background and direction. This dataset equips controlled environment to conduct studies on human activities with more stress on consistent

and repeatable actions. In contrast, UCF Sport Action gives much more details, orders of magnitude more variations in motion, and a lot of variability in the quantity. Datasets, on the other hand, are chock full of very different examples of activities in realistic backgrounds and dynamic scenarios.

The video data is pre-processed and extracted from the raw video data of KTH and UCF Sport Action datasets in an attempt to get useful features and reduce computational complexity before training the models.

Pre-processing is done, including operations such as extraction of video frames, normalization, and resizing. In the pre-processing step, we also estimate optical flow to increase the available motion information in order to render the models even more discriminative. The data augmentation techniques considered encompass random cropping, rotation, and flipping for proper sample generation during training to reduce the overfitting risk and hence achieve better generalization performance. This is to guarantee the training of the system on high-quality, representative data, which provides information properly and enables adequate learning of the intricate patterns of real-world activities in modeling scenarios. Our hope is that using this exhaustive methodology, we are able to contribute to establishing a firm basis for the accurate recognition or classification of human activities, which opens the way for new research and application in HAR.

3.1 Overview of the Models

The technical background of this research involves an in-depth understanding of methodologies related to human activity recognition and the use of transfer learning methods, including an exploration into different pre-defined models in backing up this study.

HAR has played a commanding role in understanding human activity behavior in various avenues, from health and sports analysis to surveillance. Most of the classical approaches to HAR are based on hand-crafted features from sensor data and machine learning algorithms, like support vector machines and decision trees. All these frequently encounter issues when dealing with data that is complex or high-dimensional, warranting the need for more sophisticated method.

Deep learning methodologies, such as Convolutional Neural Networks and Recurrent Neural Networks, have brought about a revolution in human activity recognition. These deep learning models are so efficient that they automatically identify the relevant features from the raw data sensed from sensors or images. In particular, CNNs show better performance in object tracking, image classification, and speech recognition. Lately, an advancement called transfer learning in the subfield of machine learning is enhancing the effectiveness of deep learning techniques in HAR tasks.

In the proposed work, we describe the technical details related to the implementation of transfer learning and its application to HAR. We leverage transfer learning to fine-tune pre-trained models, which were trained on large datasets like ImageNet, for new tasks with very few labeled data samples for their training. Again, these involve faster model convergence and enhancement in generalization

ability and performance achieved for HAR.

We discuss further a variety of predefined models used in our research: VGG16, VGG19, EfficientNetV2S, Xception, a couple of kinds of models with ConvNeXt, and others. These models are entirely different in architecture, complex structure, and performance measures. They are thus, properly considered with respect to accuracy in comparison with computational efficiency and suitability as candidates for use in transfer learning applications. This paper aims to dive into the technical underpinnings of such HAR methodologies and transfer learning techniques in the hope that it provides valuable knowledge on the effectiveness of predefined models for human activity recognition tasks.

3.1.1 Traditional Transfer Learning Models

Conventional pre-trained models have surged computer vision and are behind building most of the modern deep learning architectures today. Some standouts in this category are VGG16, VGG19, EfficientNetV2S, and Xception.

VGG16 and VGG19 are Oxford University Visual Geometry Group architectures. Models in this family have one thing in common: being very deep in architecture, with a lot of compositions of convolutional layers followed by fully connected layers. Specifically, VGG16 has 16 weight layers and VGG19 deepens the architecture to 19 layers. Although very simple in nature compared to the design of much new architecture, the family of VGG models showed impressive performance on a wide range of computer vision problems, including image classification and object detection.

EfficientNetV2S is one in the family of the state-of-the-art models, constructed with an eye towards state-of-the-art performance, while receiving equal attention to computational efficiency. Developed at Google Research, EfficientNetV2 contains a newly introduced compound-scaling approach in the balancing of model depth, width, and resolution for optimal performance on diverse tasks. In particular, EfficientNetV2S is small in size and high in accuracy, which makes it eminently suited for resource-constrained settings.

Another very popular, widely used family of pre-trained models is developed by Google. Beginning with the Inception architecture, Xception is an extension that adds something known as depthwise separable convolutions. That actually decouples the spatial and channel-wise convolution to make it computationally less complex but better in terms of representing features. So, in essence, Xception is an architecture that can capture really fine features and, at the same time, be computationally efficient. It is thus quite useful in a practical sense for a wide range of computer vision tasks.

VGG16

In the domain of transfer learning, one can site the model VGG16-pretrained. A Convolutional Neural Network architecture developed by the Visual Geometry Group of Oxford University, VGG16 has received a lot of accolades due to its

depth and performance in computer vision applications, especially image categorization.

Transfer learning may be strongly supported when the VGG16 model learns very rich features and representations from huge data. Taking into account the possible effective dealing with new challenges in classification and very few presented labeled data instances, those pre-trained weights of such VGG16 models make this very much possible. Not only does the transfer learning approach in training advance convergence speed, but it also enhances generalization and brings about high accuracy compared to training the model from scratch.

Two of the most discussed performance measures when assessing the VGG16 architecture are the Top-1 accuracy and the Top-5 accuracy. In words, Top-1 accuracy means the number of test images which were correctly classified with the highest confidence, while Top-5 accuracy means the number of test images in which the correct class lies among the top five classes forecast.

Empirical tests of the VGG16 model present state-of-the-art performance on various datasets and tasks. For instance, VGG16 shows leading results with Top-1 accuracy at 71.3% and a Top-5 accuracy of 90.1%. These two metrics underline how effective and robust the model is for delivering precise predictions as per real applications.

This is a great example of transfer learning applied in the VGG16 pretrained model: it improves performance to speed up the development of solid, very solid, and accurate classification models. Being able to learn on an intricate level, this model is suitable for researchers and practitioners involved in tasks that are very hard and limited by labeled data.

VGG19

Pretrained model VGG19 is another key ingredient in the world of transfer learning, and it is just as strong in influence and usage in the host of diversified computer vision applications. Developed by the Visual Geometry Group at the University of Oxford, VGG19 shares its lineage with the VGG16 model and has found remarkable success in a variety of tasks, most prominently in the task of image categorization.

One of the strong models is VGG19, well equipped with 19 layers to help in capturing visual data's very fine, intricate, and complex patterns. Weighing in at 549 MB, with a whopping 143.7 million parameters, the VGG19 is supercharged for learning discriminative features and achieving improved classification accuracy. Fine-tuning such a pre-trained model, the VGG19 one in specific, not only allows fast and efficient dealing with new classification problems by researchers and practitioners but makes such transfer learning possible to use learned features and representations in it for faster convergence, improved generalization, and higher classification accuracy compared with building from scratch.

Empirical studies concerning VGG19 have shown that they are very efficient and reliable in inference, although a bit lower than VGG16 in terms of performance related to the inference time. For example, tests in one paper led to an average inference time of 4.4 ms per step on a GPU and 84.8 ms on a CPU. Thus,

we can say that until today, VGG19 is still good for prudently precise predictions, considering trade-offs among inference time and model complexity when taking off. The VGG19 pretrained model is just a beacon of transfer learning powerhouse, paving the way for the improvement of performance in computer vision, more particularly in the classification of images. Its massive architecture coupled with the ability of nuanced feature extraction makes this a more pressing tool to use by researchers and practitioners in processing complicated classification tasks with limited labeled data.

EfficientNetV2S

EfficientNetV2S is heading transfer learning and picking up the curiosity of the academic research community well in its strength and compactness. Among these families of impressive models, EfficientNet2S represents the ideal regarding high performance and low demand in terms of computation and number of parameters.

Of course importantly, model EfficientNetV2S is compact at a mere 88 MB. It has shown strong performance not only in image classification tasks but also many others. This is an architecture with a balance between performance and size that infers the use of an efficient strategy to scale models.

Overall, the compact design of EfficientNetV2S would perform a better task at any instance where computational resources are constrained without a significant reduction in classification accuracy. These efficient architectural components are combined with effective scaling strategies to bring in balance between accuracy and computational efficiency, making this model a very strong candidate for any range of transfer learning applications.

EfficientNet is applied in practical aspects, especially when deploying in resource-constrained environments and real-time inferencing, capable of providing top performance while being frugal with computational resources. Lightweight, as it is, EfficientNetV2S delivers on its promise for accuracy and assured performance on every image classification task.

The EfficientNetV2S is the epitome of efficiency in the transfer learning paradigm, compact in size, high in performance, and economical in using resources. In summary, this compact yet highly powerful architecture will prove an asset for researchers and practitioners intending to apply transfer learning in resource-scarce scenarios, hence further solidifying it as a potent contender in the landscape of efficient and impactful deep learning models.

Xception

Indeed, Xception has been one of the first to emerge in transfer learning, both effective and performant, in numerous tasks associated with image classification. Famous for its small size and powerful architecture, the 88-MB Xception model has been one that many take extremely seriously among a number of circles, courtesy of its elevated precision and efficiency.

In addition, Xception has a large number of parameter powers of up to 22.9 million, relative to its small size, that makes it be able to learn detailed features

and representations from the available data. One of the striking features of this network is depth: being 81 layers deep, the network is able to acquire hierarchical representations of images and it becomes possible for the model to capture characteristics at a low level and at high levels rather accurately.

Xception makes sure that the computation is optimized with depthwise separable convolutions and the usage of skip connections. The use of such novel architectural elements places Xception at the very forefront of the realm of efficient deep learning models that strike the outstanding balance between parameter utilization and performance.

While Xception is a little slower in execution inference time in comparison with its alternatives, it is sufficiently accurate and parameter-efficient, making it suitable for serving many applications. With such times—around 109.4 milliseconds on a CPU and 8.1 milliseconds on a GPU—Xception is still relevant to problems that require real-time processing or have less strict latency requirements.

One such example that embodies the ideas of compactness, accuracy, and computational efficiency in transfer learning is shown with Xception. It can learn intricate features while optimizing the use of computational resources, thus being a valuable asset in deploying robust and efficient image classification models for researchers and practitioners in numerous real-world scenarios. Due to its relatively longer inference time, Xception becomes one of the prime choices within this setting, where the application is mainly concerned with bringing out accuracy and efficiency.

3.1.2 Introduction to ConvNext

ConvNext is working on the cutting edge to develop a structure for Convolutional Neural Networks, and new architectural elements have been incorporated to add efficiency and functionality. The ConvNet structurally differs from the other conventional ConvNets because it has multiple parallel channels, each having its configuration of convolutional layers and filters.

To use the outputs from these parallel branches, the innovative design of ConvNext concatenates them to make the final prediction. The integration with multiple branches can enable the network to distinguish intricate patterns and characteristics in input data for its effective work in tasks concerning image and speech recognition.

One big advantage of ConvNext is the extraction at both the coarse- and fine-grain level due to multiple parallel routes with different filter sizes. This should allow ConvNext to capture the widest possible variation in features, thereby allowing for greater robustness and accuracy in feature extraction.

ConvNext also showed very flexible input size configuration, a thing that classical ConvNets lacked, since spatial pooling layers downsample their input before further processing. This way, the model deals with variable input dimensions in real-world applications, which can be quite unpredictable. Moreover, the ConvNext architecture has skip connections across the input and output levels of

the network. This helps in the flow of gradient during training as well as in retaining important information from the input. This very beneficial path through input to output space obeys faster convergence under training, adding up better performance by the network.

With the addition of further novel architectural elements, ConvNext is a new landmark in Convolutional Neural Network design, which shall offer improved functionality, adaptability, and performance. It paints in the strongest and most versatile manner the possibility of this application in capturing parallel channels to take different input size optimizations under gradient flow for image recognition and natural language processing[13].

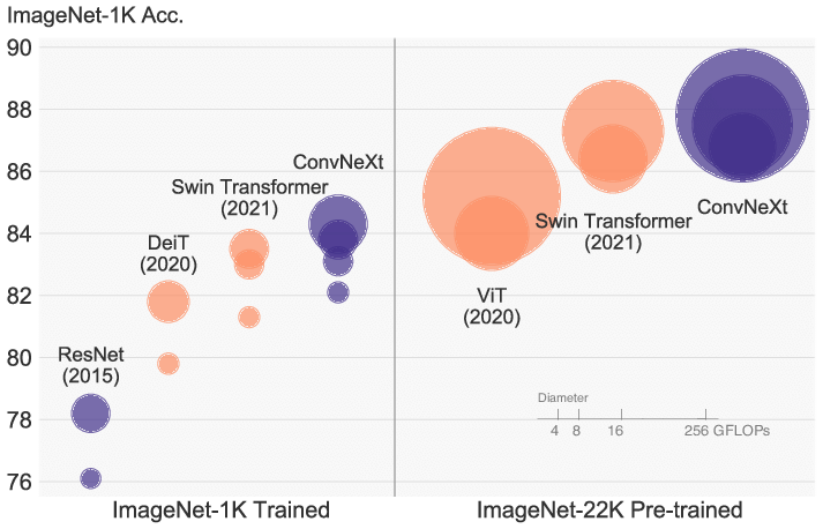

Figure 3.1: Accuracy of Pretrained Models in ImageNet Dataset

ConvNeXtTiny

ConvNeXtTiny is one of the great families of compact-sized architectures, but they do not compromise much in the native and transfer performance learning skills. It has an extremely modest footprint of 109.42 MB, yet showing tremendous capabilities to properly categorize images: it holds an accuracy rate of 81.3

Built with the idea to strike a balance between model size and functionality, ConvNeXtTiny is designed to leverage 28.6M parameters in order to successfully train discriminative features pertaining to an input dataset. The combination of cardinality (grouped) convolutions with depth-wise separable convolutions captures both spatial and channel-wise dependencies in the data, which makes the model more capable of distinguishing complex patterns and features.

Equally important, ConvNeXtTiny has a high level of accuracy compared to big models —this shows how high the capability of the model itself is in extracting useful information from the input data despite its small size. Furthermore, it has a small size of representation, meaning efficient memory usage and friendly deployment on devices with constrained resources.

This confers a lot of value to ConvNeXtTiny in relation to transfer learning across diverse datasets and domains. In this regard, researchers and practitioners

can leverage the capability of the pre-trained model of ConvNeXtTiny with applications toward image classification problems, more so when small amounts of labeled data are available. This way, quick convergence is achieved from scratch using the pre-trained weights.

ConvNeXtSmall

ConvNeXtSmall continues from the ConvNeXt Tiny family, giving a linear increase in capabilities and performance with little to no compromise in efficiency. Model size is 192.29 MB, with 82.3% for Top-1 accuracy. This is by very far the superiorly accurate, powerful, larger model in image classification compared to its predecessor.

ConvNeXtSmall uses near-similar architecture to ConvNeXt, but this time pushes the parameter numbers to 50.2M. This augmentation allows the model to extract many complex and discriminative features from the input data, making it better at recognizing subtle patterns and characteristics.

Therefore, this serves to imply that with the use of depthwise separable convolutions and cardinality (grouped) convolutions, the feature extraction from data is well improvable by ConvNeXtSmall. Although slightly larger than the former, the ConvNeXtSmall maintains the same level of efficiency in parameter utilization as is observed in the ConvNeXtTiny, thus maintaining a high level of accuracy yet accommodating the more complex representation of the data.

Practitioners will have to use a pre-trained Conv-NeXt-Small and fine-tuned on its own datasets and domain so that known features, allowing fast convergence, provide good performance, specifically for target tasks. The fact that it is adaptive and effective enough makes it the proper candidate for usage in transfer learning tasks, particularly in scenarios with sparse labeled data.

However, even if ConvNeXtSmall outperforms ConvNeXtTiny by a large margin of performance figures, practitioners need to understand the trade-off between model size and performance. Complex details are captured with more precision using ConvNeXtSmall, but in doing so, it uses more computational resources. Therefore, a careful choice has alternative ways for the requirement of the task at hand.

ConvNeXtBase

ConvNeXtBase is another big step in the architecture of ConvNeXt, oriented to deal with more complex visual recognition tasks. With enormous numbers of parameters up to 88.5 million, the ConvNeXtBase outperforms its precursors not only by complexity but also by accuracy, reaching the terrific top-1 accuracy of 85.3

ConvNeXtBase follows the previous ones in design: a chain of convolutional and pooling layers followed by one fully connected layer. However, it differs in that this model consists of more filters in each block and 'blocks' that have been added of grouped convolutions. Further, ConvNeXtBase follows the deeper superstructure by accommodating several more layers and greater kernel sizes to

assist in getting out minute patterns or features in an efficient manner from the data.

The larger size and increased complexity of ConvNeXtBase make it endowed to perform effectively in challenging tasks, showing superior visual recognition performance on tough tasks. At the same time, this ability comes at the expense of extensive training and computational resources, justifying the emphasis on resource constraints that are much needed in this regard.

In this regard, high modularity and adaptability remain at a high level, and its application in transfer learning further instills a lot of trust. The network is endowed with a feature through which researchers and practitioners can fine-tune it with their own data using pre-trained weights from ConvNeXtBase; therefore, state-of-the-art results are possible with relatively small amounts of data and computational resources.

ConvNeXtBase is the major upgrade of the ConvNeXt family, with significantly larger complexity and precision for the hardest tasks in visual recognition. Its design as a modular one, along with pre-trained weights, makes this network universally applicable to a wide range of transfer learning applications, thus advancing further the state-of-the-art efficiency and effectiveness for researchers and practitioners.

ConvNeXtLarge

Being the richest and most complicated architecture among the family of ConvNeXt, ConvNeXtLarge brings the new performance record, being extremely accurate and robust. While the model size is 755.07 MB, with the top-1 accuracy, being rated at 86.3%, the ConvNeXtLarge surpasses all its precursors and proves to be capable of being the leader in difficult visual recognition tasks.

Developed with a much higher level of complexity and parameter count—reaching 197.7 million—ConvNeXtLarge enables the extraction of complex and subtle characteristics from input data. This bigger power is due to the fact that the present design uses a higher number of filters, more blocks for grouped convolutions, and deeper networks with a higher count of layers to build up double the basic layout used for ConvNeXt.

These architectural enhancements enable ConvNeXtLarge to conduct highly discriminative representation that leads to better performance when tested with a variety of challenging visual recognition tests; however, it does so with a high computational necessity and training time, rendering it sensitive to resource constraints.

This implies that it sustains efficiency and flexibility in the models under the ConvNeXt family, even if it is scaled-up in size and complexity, hence capable of undertaking multiple transfer learning functionalities. Those pre-trained ConvNeXtLarge weights enable fine-tuning on ConvNeXtLarge with less labeled data to reach quick convergence; thus, being it attains the top performance over most areas.

ConvNeXtLarge is the major member developed within the family of ConvNeXt: the state-of-art module with high capacity in accuracy and robustness

for dealing with challenging recognition tasks. This is so great at doing service to transfer learning applications that it's a useful resource for state-of-the-art image classification models, both researchers and practitioners, to call upon for real-world scenarios.

ConvNeXtXLarge

ConvNeXt is the largest and most performant member in the ConvNext architecture family. Our large ConvNeXt, with a model size of 1310 MB and an outstanding top-1 accuracy of 86.7%, outperforms all other members of the ConvNeXt family by large margins in both scale and performance.

The state-of-the-art performance of ConvNeXtXLarge is based on the huge parameter count—350.1 million; that is what makes it capable of encoding very complex and discriminative features from input data. Relying on the base ideas of ConvNeXt, ConvNeXtXLarge introduces even more intricate and sophisticated configurations in terms of increased numbers of grouped convolutions, filters, and network depth.

ConvNeXtXLarge, while still performing excellently well, is also computationally expensive and extremely time-consuming during training because of its huge and intricate size. Accordingly, this very high accuracy places it on good footing, and it is going to be really challenged in tasks of visual recognition that are sensitive in demanding at the very highest level of accuracy and robustness.

In this context of transfer learning, ConvNeXtXLarge would remove huge obstacles: by using the pre-trained weights of ConvNeXtXLarge, one could adapt this model to new datasets and domains to utilize the learned features for performance improvement, even with small amounts of labeled data. That said, it would be very important to make some trade-offs with ConvNeXtXLarge in various aspects, such as the size of the model, computational requirements, and the possible inference time.

The ConvNeXtXLarge model is a robust and solid tool to study the potential for transferring learning and, in turn, opening up possibilities to realize state-of-the-art results on very challenging visual recognition tasks. Its superior performance and adaptability have provided this tool as a great help in the realization of up-to-the-edge solutions in image recognition and classification tasks.

3.1.3 ConvNext over traditional Convnet

ConvNext is a cutting-edge design that enhances Convolutional Neural Networks' (ConvNets') capabilities by including many parallel routes into the network.

Multiple Parallel Paths

ConvNext differs a lot from normal ConvNet because this architecture has many parallel paths. The paths have a specific number, which constitutes its convolution layers and filters, in which case it carries out its activity independently of the others. Concatenation of each individual prediction made by different paths is

what produces the final prediction. The architecture of ConvNext that can help analyze complex input information is configurable to capture diverse features and patterns simultaneously.

Enhanced Feature Extraction

ConvNext can extract both fine- and coarse-grained features from the input data since it has many parallel routes. Therefore, it can extract fine- and coarse-grained features of the input data, in many parallel routes. It really helps further in jobs that require high-level feature extraction, such as voice and picture recognition. Capturing features at various granularity levels allows the ConvNext to assimilate increasingly complex and subtle information to enhance the classification and prediction accuracy.

Adaptability to Varying Input Sizes

ConvNext uses spatial pooling layers to handle inputs of various sizes, in contrast to classic ConvNets, which generally need a fixed input size. On the other hand, spatial pooling layers are used so as to make ConvNext able to deal with inputs of arbitrary sizes. In general, classic ConvNets need a fixed input size while this is what is powerfully called by various applications in the real world where most of the input sizes for any particular task are likely to be variable, such as the processing of photographs at different resolutions. ConvNext can successfully handle inputs coming with arbitrary sizes by integrating spatial-pooling layers necessary to downsample an input before it is passed through the network.

Skip Connections for Improved Training

Skip connections, which provide direct links between the input and output layers of the network, are a feature of ConvNext. There is one important feature of ConvNext: it has direct skip connections from the input layer to the output. The result may save significant input data. These links allow the flow of training to be better due to gradients, and as a whole, less time is consumed in the process of training.

3.2 Proposed Work

It is designed to accomplish performance revolution on the human activity recognition task using ConvNeXt architectures with traditional pre-trained models in order to enhance accuracy and efficiency in classification of human activities from sensor data. Based on this, a performance comparison between ConvNeXt architecture and the performances of some traditional pre-trained models in HAR is done. More specifically, the research explores the ways in which ConvNeXt architectures function in capturing complex patterns and features in-belts with human activities and their performance compared to that of some large models, e.g., VGG16, VGG19, and Xception, in terms of accuracy and efficiency. The

study will also examine the improvement of inductive HAR performance through ensemble methods and advanced RNN architectures in connection with the ConvNeXt architectures. It will include data collection methodology—such as those from wearable devices or other sources that can be used in creating a training and evaluation set, model development through deep learning frameworks, training and validation of these models, hyperparameter tuning, and evaluation metrics, including accuracy, precision, recall, and F1-score. The experiments to be conducted involve dataset selection, setting up the experiment, and model training and evaluation with the perspective of model accuracy and efficiency compared to robustness.

Contributions should be expected to advance the state-of-the-art in HAR, acting as evidence for the effectiveness of ConvNeXt architectures, whose strengths and limitations will be comprehensively explored and further advanced toward better accuracy, efficiency, and robustness in HAR systems for real-world applications in healthcare, fitness tracking, and smart environments. Detailed timeline and milestones ensure timely completion of research objectives, including detailed data collection, development of models, experimentation, evaluation, and write-up of the thesis. In summary, the proposed work would add value to the direction HAR research is heading and contribute to the development of more accurate and efficient HAR systems.

The flowchart below describes the step-by-step process in developing and implementing a Human Activity Recognition (HAR) model by using deep learning techniques. The first step is initializing basic data structures that are created for the image data and labels, creating a list of the dataset's defined length.3.2

At this point, during the phase labeled Loading and Pre-Processing of Images, the images are loaded from the dataset, resized into 160 x 160 pixels, transformed to NumPy Array, and added to the list img_data. Its corresponding label is also added to another list called img_label. The labels are first transformed to numeric value by encoding, then redefined as one-hot encoded vector lists for machine learning model processing in the next step, pre-processing labels.

This part of the notebook builds the model. It initializes a Sequential model with Keras, loading a pre-trained model while setting the proper input shape and specifying the number of output classes. These flatten output from convolutional layers and then add a fully-connected Dense layer with 512 units, culminating in a Dense layer with 15 units.

The next stage after that would be to Compile the Model, where the model is prepared for training with a specified optimizer, loss function, and evaluation metric. Now our optimizer of choice is 'adam', the loss used is 'categorical cross-entropy', and we'll evaluate based on accuracy. In Model Training, the model is trained using the fit() method, where you pass your training data, labels, number of epochs, and your validation data. It is here, in this step, that the model weights are tuned to minimize the losses and achieve correct outcomes; finally, it is a trained HAR model. This detailed flowchart with accompanying description depicts the steps to carry out in developing an HAR model, starting from data processing to the final step: the model training part.

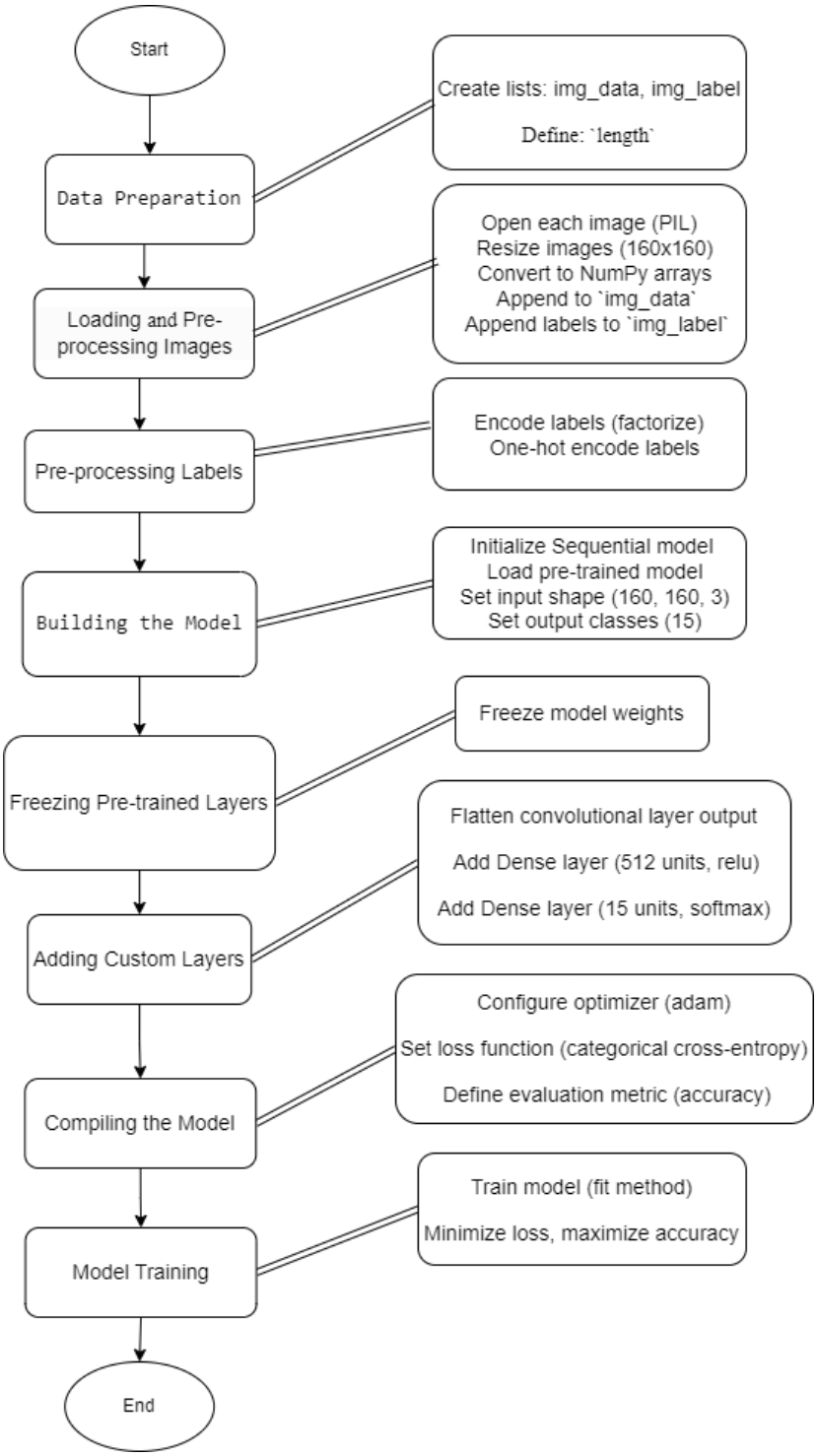

Figure 3.2: Flow chart of proposed system

Chapter 4

EXPERIMENTAL SETUP

In this section, we will discuss the implementation of different pre-trained models in HAR dataset.

4.1 About the Libraries

The following libraries provide essential functionalities and tools for data handling, model construction, training, and evaluation in the implementation.

4.1.1 Os

The os library gives the handling of datasets in a basic way, which is required to maneuver directories, manipulate file paths, and perform file operations within context. As a relevant tool in the field, os enables dynamic traversing of directories holding large configurations of image files for obtaining preprocessed data subsets for analysis. OS must support making dynamic file path creation so that on the fly it can organize data into sub-folders, based on the class labels or any other parameter, thus making datasets effectively organized and retrievable easily. Not only does it behave uniformly over any operating system but also its platform-independent, and the code is portable and robust. Having these mechanisms allows for error handling that will permit the researcher to predict and thus adjust for eventual errors regarding missing files and permission errors and, with it, increase reliability and overall robustness of the dataset management pipeline in human activity recognition efforts.

4.1.2 Glob

The ‘glob‘ library is powerful and quite fast in finding and accessing files or path names according to specified patterns, hence allowing for a very imperative way of completing data management tasks within your project. This library enables the easy dynamic loading of image files or datasets that have been stored in different directories, hence reducing many efforts during data preparation before training or evaluation. This flexibility will be assisted in merging data from

diverse sources seamlessly and also allows one to build up large datasets for tasks such as human activity recognition. ‘glob‘ also allows the researcher to input specified search patterns needed, like wildcard expressions, so as to gain an extra level of granularity and get to a higher level of detail in the identification of files or directories. In general, ‘glob‘ enhances productivity and data quality in the context of management workflows, which enables its users to focus on creating and assessing trustworthy human activity recognition models.

4.1.3 Numpy

The ‘numpy‘ package is a mainstream library in the numerical computing environment, with efficient functions for working over arrays, mathematical transformations, and general data manipulations. It supports a lot—just a lot—of tasks for analysis and model formulation pertaining to human activity recognition. It enables one to perform mathematical computations on image data effectively, such as normalization, scaling, and dimensionality reduction. This is followed by responsible data preparation for further analysis. Furthermore, ‘numpy‘ supports the extraction of salient features from raw data for the generation of informative representations that encapsulate major characteristics of human activities. Moreover, ‘numpy‘ can yield input tensors that fit perfectly into deep learning frameworks, so compiling and evaluating models is much easier. By utilizing the power of the ‘numpy‘ library, new knowledge from raw data can be elicited in order to develop human activity recognition systems that are more accurate and efficient.

4.1.4 Pandas

Other libraries, such as ‘pandas‘, are flexible enough and, further, very important in terms of working with and analyzing data. What it specifically has to deal with is the structured-data formats like CSV files. Meanwhile, in the project you are engaged in regarding human activity recognition, there will be provided an extended number of tools for effective table data management with information about different activities. Through intuitive functionalities, a user is capable of loading CSV files containing activity labels, merging datasets for compactness, and executing data transformation, changing them into a form that will be easy to proceed with further into analyses or model training. Similarly, ‘pandas‘ allows descriptive statistics to be extracted, which are important from the point of obtaining relevant information about the basic characteristics of the activity and distributions:. With its intuitive structure and huge functionality, ‘pandas‘ makes it easy for a scientist to help simplify the data preprocessing pipeline and speed up the process of the base model development to recognize activities.

4.1.5 Tensorflow_Addons

‘tensorflow_addons‘ is a good wrapper over TensorFlow, incorporating many features alongside specialized operations and utilities. For human activity recogni-

tion, in particular, ‘tensorflow_addons‘ has been designed with advanced capabilities, including specialized layers, which work hand in hand with the core functioning of TensorFlow. These include a specialized loss function to handle some special aspects of activity recognition, specific metrics to measure model performance, and complex layers that capture the many aspects of human activity data properly. Inclusion of ‘tensorflo_addons‘ into your workflow will put researchers in a better position to have access to many tools and resources through which it will rather be easy to develop very high-performance activity recognition models. This seamless integration allows researchers to explore new and innovative ways for the optimization of model architectures so as to develop competitions that rely on the best performance for accurate and robust recognition.

4.1.6 Tensorflow

TensorFlow forms the very basis of deep learning frameworks that provide a full set of tools and functionalities in creating, training, and deploying machine learning models. You will use TensorFlow as the foundational framework in building and training neural networks in your human activity recognition project. With the rich ecosystem of TensorFlow, researchers will benefit from a wide array of tools, APIs, and prebuilt models for developing advanced architectures, particularly crafted for problems as interesting as activity recognition. The flexibility of TensorFlow allows working with different paradigms in deep learning: from ConvNets, RNNs to transformer-based architectures, hence leaving where new ideas or experiments in model design can be brought in. This is where TensorFlow can be quite handy, with powerful functionalities applied to make the development process easy for the researchers and further optimize model performance toward state-of-the-art results for human activity recognition.

4.1.7 Keras

Keras is an integral part lying in close proximity to TensorFlow and is a high-level API of neural networks. It is designed to be fast, user-friendly, and easy to stick together for researchers with small code complexity during model building, composing, training, and evaluation. In the context of your project, Keras eases the process of designing and putting into implementation architectures relevant to deep learning during human-activity-recognition tasks. Due to the user-friendliness and modular architecture of Keras, it is very easy for researchers to design and configure neural network models according to the subtleties of activity recognition tasks. There are a variety of pre-configured layers, activation functions, and optimization algorithms in Keras for fast prototyping and research on the configuration of a model. Additionally, Keras has very good integration with the TensorFlow ecosystem. It further enables smooth interoperability by allowing the use of a rich collection of other tools and resources for the training and evaluation of models in this ecosystem. In service to this overall goal, the user-friendly design and wide functional coverage of Keras allow the research activities needed

to speed up the development cycle in order to optimize model performance for attaining cutting-edge results in activity recognition.

4.1.8 Layers (by Keras)

There are some pre-configured layers available inside Keras's 'layers' module that could be helpful in building a neural network. They span from a large list of convolutional, pooling, activation, and normalization operation layers—basic building blocks for the composition of deep neural networks. For your project, it will help in predefining layers for the optimal development and configuration of a Convolutional Neural Network (CNN) or a Recurrent Neural Network (RNN) model, specifically for recognition during the observation of human activity. Such activation layers will allow researchers to build deep architectures very easily and experiment with many settings in tuning parameters toward better agreement in performance.

4.1.9 ImageDataGenerator

It gives the 'ImageDataGenerator' tool, and in return, it acts as a powerful resource for real-time data augmentation during the training period of an image dataset. A case in point for human activity recognition: with the usage of the 'ImageDataGenerator', one can augment their data by applying transformations such as rotation, scaling, shifts, and flipping on input images. This kind of strategy in increasing the dataset's richness with respect to diversity of training samples boosts the robustness and generalization capabilities of deep learning models. Using an 'ImageDataGenerator', you can greatly reduce overfitting and improve the performance of a model on new data. This is done by subjecting the model to greater spectra of variability that might be present in the original dataset.

4.1.10 Categorical

The 'to_categorical' function is very important since it converts class vectors to one-hot encodings. This operation is done mostly in classification that uses categorical labels. In your project for human activity recognition, class labels denoting various kinds of human activities are now in the form that is suitable for training the neural network. It means by converting this, the output of the model can be interpreted as probabilities across classes, making it ideal for an exact prediction and evaluation of an activity recognition task. By using 'to_categorical', you are sure that your model will learn correctly to classify activities with respect to the categorical representations thereof, which is more effective and interpretable.

4.2 Dataset

A well-curated and annotated dataset of images provides the backbone to accurately label these images in this research work. This supposedly encompasses something like 15 different categories, which people generally divide human activities into, and special care was taken for these categories to crop up in very different kinds of actions and movements: walking, running, sitting, standing, jumping, or for that matter, more specialized activities such as dancing, practicing, playing different sports, etc [5]. This project aims at creating a diverse and representative dataset, which could be used not only in training but also in cross-validating machine learning models applied for Human Activity Recognition.

The data set consists of nearly twelve thousand images, divided into training and validation sets for ease of building a model and evaluating it. Each image was perfectly described, characterized by an easy-to-observe name that attributes the activity being undertaken. This is required to be foundational with respect to supervised learning: that it be the ground truth for which you train the models and they learn to recognize and classify activities.

The images used in the selection process have to reflect a wide variety concerning every category but also of every different context and condition of the images: lighting, background, subject appearance, movement dynamics, and so on. This is very important for building a model that is strong in generalization toward real-world situations. The images were collected from various environments, being both outdoor and indoor. This was in an effort to represent a wide scope of potential conditions under which the activities could take place. Furthermore, preprocessing techniques were applied for clarity and relevance grooming of each of the images: steps applied prior to cropping, resizing, and normalization of the input data, making it so that it could be standardized with the ML algorithms. Additional techniques to filter out noise totally cleaned these images of irrelevant detail that would only hamper the learning process.

It is well equipped with the all-inclusive information in a dataset; hence, models developed from it return not only accurate results but also reliability in identifying and classifying a broad spectrum of human activities under different scenarios. This type of rich dataset is therefore very important in training deep learning models because it incepts the model with a lot of diversity in examples to learn from and therefore develops it to be able to recognize subtle variations in the activity being conducted. The approach aims to build models for correct HAR with high accuracies and robust frameworks, relying on a wide and varied dataset. Accordingly, it trains models that predict by a classification of the correct activity class which the image it contains is described by its label. Generalization performance of the built models has been measured with validation on new unseen data.

Being a diverse data set of almost 12,000 labeled images, this well-structured collection is one strong base for ensuring further development towards the state-of-the-art HAR models. The careful selection, grooming, and labelling of the dataset assure that models trained on the data can perceive a wide class of activities with

high accuracy and confidence, thus making strong contributions to the field of activity recognition and its practical application.

The fifteen human activity categories included in the dataset are as follows:

1. Calling
2. Clapping
3. Cycling
4. Dancing
5. Drinking
6. Eating
7. Fighting
8. Hugging
9. Laughing
10. Listening to music
11. Running
12. Sitting
13. Sleeping
14. Texting
15. Using a laptop

To provide a visual representation of the dataset's diversity, the image (Figure 3.1) below showcases samples from each of the fifteen activity classes, illustrating the breadth and richness of human activities captured within the dataset.

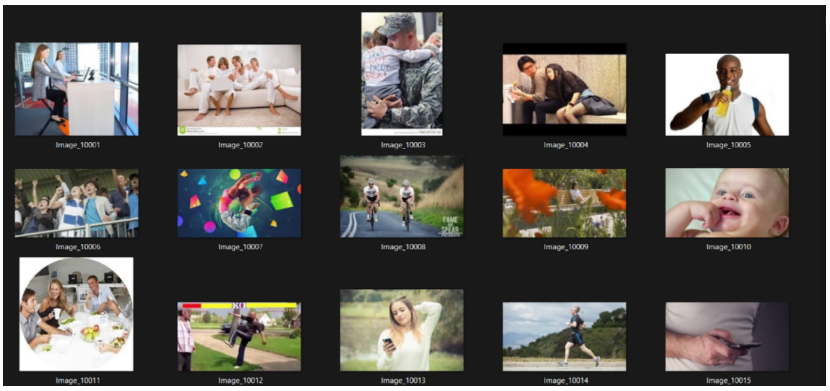

Figure 4.1: Human Activity Images taken in Dataset

Besides associating the dataset, which is broad in coverage with respect to human activities related to different stages of life, the annotated nature of datasets really proves to be a valuable resource both in developing and evaluating any activity recognition model. Every image in the dataset is annotated well enough through a descriptive tag representing an activity coherently and truthfully; this lays down a nice foundation for the supervised learning processes. This detailed annotation is very crucial in order to teach models to recognize, classify, and learn activities based on human movement in a correct way based on good quality labeled input samples.

The dataset has a very general scope covering 15 classes of human activities, ranging from the basic daily activity of a human being to more specific ones such as dancing, exercising, and playing sports.

This ensures diversity in the data set and captures a very wide array of movements and contexts that might mirror human behavior in all its complexity. It captures activities carried out across many different environmental settings: indoor and outdoor, hence representing the nature of conditions that these activities are carried out in.

This further explains that, from around twelve thousand images, it is subdivided for training and validation purposes. The model learns from a big number of examples within the training set, whereas the purpose of validation is supposed to check performance. It should check how general the model can be to new, unseen data. Without such splitting, there is no possibility to build a really well-working model, one which may perform well under real-life circumstances.

Preprocessing is essentially structured with cropping, resizing, normalization, and noise deduction in order to obtain improved input data quality and harmonize data for machine learning algorithms, while at the same time preventing overfitting risks. Then each image is groomed carefully so that each one becomes relevant and clear to provide the best training data to the models. Leveraging the richness of annotation and diversity within this dataset, it will be possible to create highly sophisticated state-of-the-art models for human activity recognition not only in terms of accuracy but also in terms of reliability across contexts and conditions. The extensive amount of coverage on human activities in detail through the labelling of images further generalizes the dataset so that new research in human activity recognition can be initiated.

Models built on this dataset can perform well, which can be expected in health care applications, fitness tracking, and smart home systems with surveillance.

Indeed, it is a rich data set in the sense that it contains varied activities throughout the life course, and hence studying human behavior change over time would be very valuable. For example, the study will be able to investigate how recognition models can account for changes in an activity pattern through the inclusion of pictures and videos showing activities carried out by individuals at different life stages. Together, the diversity and annotated nature of the dataset, together with comprehensiveness and wide coverage of human activities, make it a major stride in human activity recognition research. This is indeed very important for providing a strong base in the development of models that take a lot of activities with accuracy and dependability through novel applications, which finally add to human behavior understanding. This dataset, with rich annotations and wide examples, will undoubtedly become very important and could shape the future related to research in activity recognition.

4.3 Data Pre-processing

Human activity recognition mainly relies on the performance and reliability of the model with very high-quality and consistent data. Proper data preprocessing is an essential step toward ensuring the quality of the dataset for training advanced models like the ConvNeXt architecture. This section elaborates on the data preprocessing pipeline followed in this work to ensure that data are problem-free,

and thereby meet the very strict standards for human activity recognition.

The philosophy that underlines our preprocessing pipeline is the class consistency and balance of all categories of activities. This principle is embedded at every step, from data collection through preprocessing, and ensures that the process of training and evaluation is run on consistent and balanced representations of each activity class. These aspects focus on the minimization of bias against some specific activity groups, which otherwise would distort the model's performance and further affect generalization.

The above is a comprehensive data preprocessing pipeline for ensuring high-quality and research-friendly HAR datasets. The main intention for the class consistency and balance in this dataset is to make the dataset fair, hence not biased since reliable and accurate models for activities have been developed. Such an approach enhances the model performance of the ConvNeXt architecture, hence making it a general objective for advancing research on human activity recognition using a robust and reliable foundation for training and model evaluation processes.

From the above sections, it is evident that the data of images along with the corresponding labels are structured in a properly arranged CSV file. The file acts as the center point for storing each image's label, so management and retrieval could be done in a simplified way in further pre-processing and training stages.

The big part of pre-processing the data is to make sure the source photographs are all a uniform size and format. All images are uniformly compressed to 160x160 pixels in resolution. Scaling ensures that all images are equal, which can be used for comparison and feature extraction when the models are being trained.

The images are then converted into numpy arrays after resizing by using the Python numpy module. Python's np.asarray() method quickly converts these images into numpy arrays so that they can be efficiently handled and processed within the structure of the ConvNext architecture. These Numpy arrays encode the preprocessed image data, thereby ready for both subsequent feature extraction and model training processes.

At the same time, the class labels defined for each image are passed through a transformation with the numpy library function "to_categorical()". At this transformation step, the class labels are encoded as one-hot vectors in order for the model to be appropriately guided by categorical nature of classes and to predict human activities properly. Such encoding of a categorical label makes the dataset compatible to the ConvNext architecture which helps in distinguishing and categorizing human actions effectively by a machine learning model.

The pre-processing procedures are important and, as a result, should be carried out in such a meticulous way that they indeed create a dataset tailor-made for use in human activity recognition. In this study, the suitability and effectiveness of the dataset have been ensured by adopting uniform picture sizes, balanced class distribution, and categorical label encoding techniques for training effective ConvNext-based activity recognition models.

4.4 Implementation Procedure

This section presents in detail the step-by-step process and flowchart of the human activity recognition model implementation using deep learning techniques. In detail, it describes prepared tasks or operations to be applied in model development, training, and evaluation, starting from stages of data preparation and pre-processing, right through to model construction, compilation, and training. From this, the steps can guide the reader across an entire workflow on how the implementation is done in a way that is easy to grasp.

This section also indicates the necessity of the libraries and functions—following are some: TensorFlow, Keras, NumPy, PIL in their respective roles throughout the implementation. The general scope of this section will be a guide for practitioners and researchers in replicating or adapting the illustrated model for their own use in a human activity recognition task.

4.4.1 Data Preparation

It's in this step that you create two lists: 'img_data' and 'img_label'. You have also defined the variable 'length', which indicates a total number of training data instances. This step initializes the most basic data structures for the loading and processing of the dataset.

4.4.2 Loading and Pre-processing Images

This is the stage where images from the dataset are loaded and pre-processed for further processing. You iterat ovr the lngth of th training data, opn ach imag with th Python Imaging Library, rsiz th imags to a fix siz with prdfind dimnsions of 160x160 pixls, convrt th imags into NumPy arrays, and appnd ths lists of NumPy arrays to th list img_data. At the same time, furthermore, the corresponding labels for each of the images are appended in the list img_label. This process is critical to bring the dataset into memory and ready for the next processing step.

4.4.3 Pre-processing Labels

At this step, you preprocess the labels that come with your images. The factorise function is applied to the labels using the train_csv DataFrame. By encoding categorical labels into numeric values, it makes it possible for a machine learning model to understand them. Then the numeric labels are converted into one-hot encoded vectors by calling the 'to_categorical()' function. One-hot encoding is a very common method in classification tasks. In this method, the vector representation of a label transforms into a binary vector where the value '1' states that one element belongs to a class and the value '0' states that it does not.

4.4.4 Building the Model

Here you will start to construct the model of a neural network for human activity recognition. First of all, this is initializing a Sequential model with Keras, which is actually a high-level neural network API. The pre-trained model is loaded, leaving only the last fully connected layers. The 'input_shape' parameter equals (160, 160, 3), meaning size of input images: width, length, and 3 for RGB. The number of output classes is also set to 15, which matches with respect to the number of activity categories in the dataset.

4.4.5 Freezing Pre-trained Layers

You are freezing the weights of the model's pre-trained layers, fixing them during training. The aim is to avoid changing or retraining the weights so that it can keep their learned representations from the pre-trained model. This way, it's less computationally expensive and guards, from a statistical point of view, against the possibility of the model forgetting features that it learned during the process of learning a new one.

4.4.6 Adding Custom Layers

Here, you add custom layers on top of the pre-trained model to tailor it to the specific task of human activity recognition. The output of the convolutional layers is flattened to prepare for the fully connected layers. Subsequently, a fully connected "Dense" layer with 512 units and a 'relu' activation function is added to capture high-level features from the extracted representations. Finally, a "Dense" layer with 15 units (equal to the number of classes) and a 'softmax' activation function is added to output probabilities for each activity category.

Going further, you tailor the model to custom requirements of your task by adding some custom layers on the top of pre-trained models. The output from convolutional layers is flattened so as to prepare the features ready for the fully connected layers. This is then followed by a fully connected 'Dense' layer with 512 units and an activation function of 'relu' in order to capture high-level features from the extracted representations.

Finally, we use a "Dense" layer that has 15 units with the same number of classes to obtain the category probabilities, where 'softmax' is used as the activation function.

4.4.7 Compiling the Model

In this step you compile the model in order to configure it for training. The 'compile()' method is used in specifying the optimizer, loss function and evaluation metric. Here, the 'adam optimizer, 'categorical cross-entropy' loss function along with the 'accuracy' metric is selected. The optimizer is in charge of updating the model's weights over training, and the loss Function approximates the per-

formance of the model during optimization. The accuracy metric evaluates the model's performance against the validation dataset.

4.4.8 Model Training

Then, the model is trained on the training data. The method 'fit()' is supplied with associated training data, labels, the number of epochs to train for, and validation data it should check against. The model is trained in a way that minimizes loss and increases accuracy by tweaking its weights through an optimization algorithm and a loss function. This argument was set to '1' so that I could obtain logs for every epoch to be used later in gaining some insights into the training process.

Chapter 5

RESULTS AND DISCUSSION

This chapter presents the training process's outcomes and the experiments in which different models were used in the context of human activity recognition. It sets deep insights into model performance measured in terms of accuracy, loss, and other relevant measures that are drawn out of training and evaluation phases. The performance variance is also compared in the classification of the exact human activity for all models. This gives readers a very clear idea of the relative performance and suitability of each model for the specified task, which can help them make decisions and further explore the results.

5.1 Training Process Results

This section shows the result of training process of all the models used in the experiment.

5.1.1 VGG16

Epochs	Train Accuracy	Train Loss	Validation Loss	Validation Accuracy
1	0.4354	2.3097	1.5611	0.5012
2	0.6115	1.2049	1.5441	0.5417
3	0.7031	0.9072	1.5237	0.5500
4	0.7813	0.6647	1.6445	0.5270
5	0.8478	0.4721	1.8451	0.5377
6	0.8979	0.3234	1.9071	0.5504
7	0.9292	0.2246	2.1260	0.5345
8	0.9571	0.1435	2.1318	0.5421
9	0.9762	0.0966	2.3576	0.5460
10	0.9558	0.1446	2.5895	0.5278

Table 5.1: Training process of VGG16

The table below summarizes the performance of the VGG16 model on HAR for 10 epochs. Validation accuracy has a slight increase, remaining around 50%, whereas training accuracy increased from 43.54% up to 95.58%, and training loss dropped to 0.1446 from 2.3097. The best validation accuracy achieved is 55.04%. It overfits because it does not generalize as well with new data as it does with the training data. This can be explained by the fact that the validation loss increases up to 2.5895.

5.1.2 VGG19

Epochs	Train Accuracy	Train Loss	Validation Loss	Validation Accuracy
1	0.4450	2.2763	1.5712	0.5044
2	0.6137	1.2110	1.5558	0.5222
3	0.7059	0.9099	1.5755	0.5417
4	0.7837	0.6627	1.6346	0.5540
5	0.8437	0.4770	1.7299	0.5595
6	0.9020	0.3130	1.7934	0.5480
7	0.9350	0.2150	1.9439	0.5603
8	0.9658	0.1336	2.0790	0.5540
9	0.9788	0.0905	2.2085	0.5571
10	0.9823	0.0722	2.3624	0.5591

Table 5.2: Training process of VGG19

Training for the Human Activity Recognition VGG19 model happens over 10 epochs. The training started from 44.50% accuracy and 2.2763 loss in epoch 1; then, it improved to 98.23% accuracy with a 0.0722 loss in epoch 10. Validation accuracy increases just a bit at the end of epoch 0, from 50.44% to 55.91%, signifying that there are some generalization issues. Over the epochs, there is evidence of a trend in the increase in validation loss which has a maximum at 2.3624. These trends suggest potential overfitting and require further development in optimization processes to better model generalization on unseen data.

5.1.3 EfficientNetV2S

Below is training performed on EfficientNetV2S for 10 epochs in the case of Human Activity Recognition. Starting from an initial training accuracy of 60.62 with a corresponding loss of 1.2529 in epoch 1, the two metrics continuously rise across the training. At the 10th epoch, the model hits a training accuracy of 94.41 along with the loss of 0.1800. The validation metrics are also very telling because the accuracy surges from 65.60 to 67.42% over the same period, yet the loss slightly arises at 1.3200. Such results indicate an effective learning process and the possibility of generalization, although there is room for slight optimization.

Epochs	Train Accuracy	Train Loss	Validation Loss	Validation Accuracy
1	0.6062	1.2529	1.0638	0.6560
2	0.7137	0.8851	1.0497	0.6611
3	0.7616	0.7311	1.0305	0.6655
4	0.8137	0.5886	1.0178	0.6829
5	0.8431	0.4798	1.0695	0.6746
6	0.8826	0.3713	1.1150	0.6734
7	0.9018	0.3040	1.1518	0.6714
8	0.9253	0.2421	1.2077	0.6734
9	0.9342	0.2118	1.2540	0.6698
10	0.9441	0.1800	1.3200	0.6742

Table 5.3: Training process of EfficientNetV2S

5.1.4 Xception

Epochs	Train Accuracy	Train Loss	Validation Loss	Validation Accuracy
1	0.1355	5.1623	2.5775	0.1444
2	0.1819	2.4961	2.4989	0.1698
3	0.2156	2.3857	2.4235	0.1956
4	0.2309	2.3212	2.4329	0.1901
5	0.2536	2.2518	2.4419	0.1988
6	0.2686	2.2085	2.4543	0.1988
7	0.2854	2.1506	2.3898	0.2123
8	0.3001	2.0985	2.4336	0.2028
9	0.3180	2.0535	2.5453	0.1960
10	0.3322	2.0043	2.5273	0.1972

Table 5.4: Training process of Xception

The training results for the Xception model in Human Activity Recognition are then logged over 10 epochs. From both initial metrics, it is seen that at epoch 1 are a training accuracy of 13.55% and loss equal to 5.1623. They both somewhat increase along with the continuation of the next epochs. However, for epoch 10, these remain significantly low at 33.22% for training accuracy and 2.0043 for loss. Marginal improvement in validation metrics is also seen, with the resultant accuracy of 19.72%, which is quite low. The results point out some challenges for model learning and generalization: they seem to bring about a need for further optimization and exploration.

5.1.5 ConvNeXtSmall

The training progress of the ConvNeXtSmall model is plotted over 10 epochs for Human Activity Recognition. Both starting from an initial training accuracy of 66.61%, and loss at 1.0662 in epoch 1, they slightly increase all through the training. After the 10th epoch, the model has a training accuracy of 98.44% with a loss value of 0.0569. Validation metrics further show the good trends of increase: accuracy rises to 71.94% and validation loss has gone down just a little to 1.3656. The classification results show considerable learning and generalization of the model, which may perform robustly on unseen data with further optimization.

Epochs	Train Accuracy	Train Loss	Validation Loss	Validation Accuracy
1	0.6661	1.0662	0.8929	0.7190
2	0.8107	0.5923	0.8487	0.7365
3	0.8817	0.3856	0.8845	0.7405
4	0.9235	0.2511	0.9320	0.7353
5	0.9535	0.1645	0.9602	0.7480
6	0.9517	0.1074	1.0224	0.7437
7	0.9700	0.0715	1.0696	0.7496
8	0.9864	0.0543	1.1557	0.7484
9	0.9890	0.0435	1.2181	0.7433
10	0.9844	0.0569	1.3656	0.7194

Table 5.5: Training process of ConvNeXtSmall

5.1.6 ConvNeXtBase

Epochs	Train Accuracy	Train Loss	Validation Loss	Validation Accuracy
1	0.7801	0.7238	0.5805	0.8250
2	0.9043	0.3010	0.5714	0.8333
3	0.9454	0.1749	0.6133	0.8325
4	0.9700	0.0972	0.6573	0.8357
5	0.9834	0.0604	0.7114	0.8421
6	0.9869	0.0440	0.7631	0.8353
7	0.9899	0.0341	0.7667	0.8361
8	0.9880	0.0394	0.8749	0.8290
9	0.9771	0.0786	0.9555	0.8143
10	0.9745	0.0797	0.9990	0.8163

Table 5.6: Training process of ConvNeXtBase

The training results of the ConvNeXtBase model with respect to HAR are logged through 10 epochs. In the beginning, the obtained training accuracy was at 78.01%, and the loss is 0.7238 in epoch 1, although both measures constantly rose from the following epochs. Finally, in epoch 10, the training accuracy reached 97.45% with a loss of 0.0797. Validation metrics also seemed promising, in which the accuracy increased to 81.63%, but the validation loss crept up in value to 0.9990. It indicates effective learning and generalization ability, hence showing potential for robust performance on unseen data with further optimization.

5.1.7 ConvNeXtLarge

Epochs	Train Accuracy	Train Loss	Validation Loss	Validation Accuracy
1	0.8020	0.6631	0.5230	0.8413
2	0.9228	0.2443	0.5587	0.8341
3	0.9633	0.1178	0.5659	0.8508
4	0.9803	0.0644	0.6850	0.8440
5	0.9856	0.0479	0.6796	0.8437
6	0.9921	0.0295	0.7294	0.8504
7	0.9956	0.0166	0.7515	0.8575
8	0.9960	0.0164	0.8715	0.8444
9	0.9697	0.1011	0.9436	0.8357
10	0.9636	0.1145	0.9751	0.8286

Table 5.7: Training process of ConvNeXtLarge

The illustration below shows the training progress for human activity recognition with the ConvNeXtLarge model after 10 epochs of training. Starting accuracy is 80.20% and loss is 0.6631 at epoch 1, both monotonically increasing during training, it gives a training accuracy of 96.36% and an average training loss of 0.1145 over 10 epochs. Further promising trends can be seen from validation metrics: the accuracy now reaches 82.86%, and the loss slightly rose up to 0.9751. All those results would lead us to believe that learning and generalization are effective; thus, the potential of the model for good performance with still not foreseen data is pretty solid.

5.1.8 ConvNeXtXLarge

Training progress of the ConvNeXtXLarge model for human activity recognition over 10 epochs: starting from a training accuracy of 80.55% and a loss of 0.6631 in epoch 1, development continues quite smoothly for both metrics. In epoch 10, the model receives a training accuracy of 98.97% with a loss of 0.0317. Validation metrics also show promising growth; for example, accuracy increased to 85.15%, and validation loss slightly dipped by a notable 0.9744. The results indicate a

Epochs	Train Accuracy	Train Loss	Validation Loss	Validation Accuracy
1	0.8055	0.6631	0.4954	0.8512
2	0.9213	0.2450	0.5701	0.8365
3	0.9603	0.1262	0.5908	0.8460
4	0.9806	0.0621	0.6144	0.8480
5	0.9829	0.0538	0.7148	0.8508
6	0.9854	0.0483	0.7876	0.8448
7	0.9768	0.0738	0.9707	0.8222
8	0.9745	0.0794	0.8929	0.8440
9	0.9869	0.0409	0.9527	0.8472
10	0.9897	0.0317	0.9744	0.8516

Table 5.8: Training process of ConvNeXtXLarge

good process of learning and generalization, churning out models that could serve honorably for the classification task on unseen data.

5.2 Compare between eight models on training and validation data

This sections provides the comparison of all the predefined models. Comparison

Sl.No	Pre-trained Model	Training Accuracy	Validation Accuracy
1	VGG16	97.46	56.23
2	VGG19	98.43	54.96
3	EfficientNetV2S	96.71	68.85
4	Xception	53.68	23.36
5	ConvNeXtSmall	99.86	71.80
6	ConvNeXtBase	98.13	83.41
7	ConvNeXtLarge	98.63	85.63
8	ConvNeXtXLarge	98.23	85.87

Table 5.9: Comparison of 8 models

between eight predefined models for human activity recognition in HAR amid the training and validation data.

The training accuracy of the VGG16 model is high at 97.46%, but the validation accuracy is very low, at 56.23%, which proves quite overfitted. Equally high is the training accuracy for the VGG19 model at 98.43%, with the validation one being even lesser at 54.96%, suggesting an analogous nature of overfitting.

On the other hand, the EfficientNetV2S model has slightly lower training accuracy, at 96.71%, but spikes to relatively high validation accuracy of 68.85%,

hence generalizing better. It might imply that the EfficientNetV2S model may be decoding the primary information embedded in the dataset without overfitting.

In contrast, the prediction made by the Xception model is rather poor on training and validation datasets, with accuracies of 53.68% and 23.36%, respectively. This implies that it is challenged in learning proper representations for data, hence the poor performance in both sets.

The ConvNeXt models consistently outperform the VGG and Xception models in both training and validation accuracies. An exceptionally high training accuracy of 99.86% and 71.80% for validation accuracy was achieved with the ConvNeXtSmall model, indicating robust learning and generalization capability. On the other hand, baseline models such as ConvNeXtBase, ConvNeXtLarge, and ConvNeXtXLarge also showed a lot of contrasts in terms of training and validation accuracies, which ranged from 98.13% to 98.63% for training and from 83.41%.

These results appear to support the fact that ConvNeXt models are highly effective for HAR tasks, and among them, ConvNeXtLarge and ConvNeXtXLarge particularly perform strongly. This shows the importance of model architecture and design in handling HAR tasks. A sign of how classic models like VGG and Xception perform a little limit to capturing complex patterns from data compared to that of newer architectures like EfficientNetV2S and ConvNeXt, which perform well and show high promise in HAR applications due to high accuracy and generalization abilities.

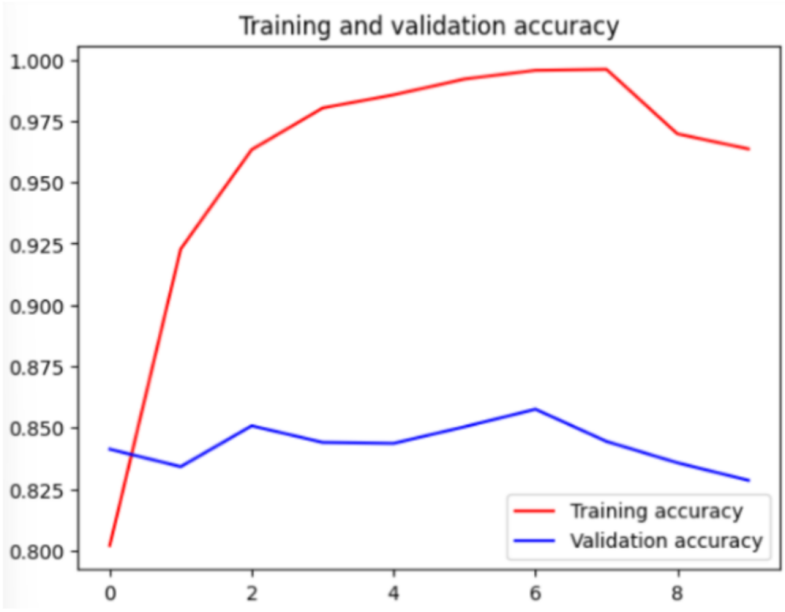

Figure 5.1: Training and validation accuracy of ConvNeXtLarge Model

Figure 5.2: Training and validation Loss of ConvNeXtLarge Model

Chapter 6

CONCLUSION

In the field of HAR, the pursuit of greater accuracy and efficiency has led the recognition community from considerations about everything from traditional pre-trained models to more advanced architectures like ConvNeXt. Here, this paper deals with the deep evaluation and comparison of their performances on a dataset comprising 12,000 images in 15 different activity classes. The results of the experiment produced an interesting story that definitively highlights the superiority of ConvNeXt architectures applied to HAR over all others, with ConvNeXtLarge leading this group. Even though promising performance initially appeared using traditional models pretrained as VGG16, VGG19, EfficientNetV2S, and Xception, their performance was largely poorer than for newly formulated ConvNeXt ones.

Of all the trained models, the leading model proved to be ConvNeXtLarge, which touched on an accuracy of training and validation better than the others. This major success has therefore proven the power of ConvNeXt architectures in capturing complicated patterns and features that are generally displayed in human activities. By using a mixed convolutional layer architecture in ways never before done, ConvNeXtLarge managed to state-of-the-art accuracy and robustness of activity recognition. It can pick out fine flavor and adapt to mixed datasets, thus putting it at the core of where HAR is moving. Moreover, success of model ConvNeXtLarge is of more general importance for the machine learning/artificial intelligence community in that this attests to the fact that new design in architecture might solve major practical tasks which claim high precision/reliability, in particular HAR.

This research also depicted how important continuous innovation and exploration are to the development of models. With the ever-evolving technologies and the increasing complexity of the datasets, new methodologies and architectures are coming forward which will now start to come into reality. The great performance of ConvNeXt Large will not only consolidate itself further among models for HAR but also open new possibilities and potential for deep learning.

6.1 Future Work

The developed ConvNeXtLarge model demonstrated excellent performance, the architecture should be further investigated and optimized. This can be done with further experimentation with different configurations, layers, and hyperparameters to increase generalization capabilities toward datasets with even more diversity and complexity in activity patterns. It will allow data augmentation and semi-supervised learning to enhance further the robustness and performance of HAR ConvNeXt models when a few labeled data are available.

Another important point is bringing in multi-modal data sources in the future. While the present study was majorly based on visual data, merging information from other sensors such as accelerometers, gyroscopes, and physiological sensors may provide a much clearer picture of human activity in general. This might call for techniques of fusion novel enough to bring heterogeneous sources of data cogently together for more accurate, context-aware HAR systems. Furthermore, such research should be applied and deployed further. This would include problems that arise due to real-time processing, computational efficiency issues, as well as scalability of the models. In terms of practical application of HAR, there is an urgent need to develop light and efficient variants of the ConvNeXt model architecture that are executable on edge and smartphone devices.

The final indication, however, is to the monitoring of the HAR systems in relation to adapting to the dynamic nature of environments. With the advent of new types of activities and changes in user behavior observed, online learning techniques and adaptive algorithms will allow an HAR system to equip itself with the tools to evolve while keeping performance high. Future works will explore the optimization of ConvNeXt architectures for joint learning from multimodal data with the help of advanced RNNs, incorporation of the application of different techniques for explainable AI, deployment in real-world scenarios, and continued adaptation. All these steps together will advance the field of HAR further toward more accurate, robust, and reliable activity recognition systems that can cater to the demands of actual and practical real-world environment settings.

Bibliography

[1] Fatimah Al Heeti and Muhammad Ilyas. Comparative analysis of convolutional neural network architectures for classification of plant leaf diseases. In *2022 2nd International Conference on Computing and Machine Intelligence (ICMI)*, pages 1–5. IEEE, 2022.

[2] Muhammad Arif and Ahmed Kattan. Physical activities monitoring using wearable acceleration sensors attached to the body. *PloS one*, 10(7):e0130851, 2015.

[3] Ahmed Ayman, Omneya Attalah, and Heba Shaban. An efficient human activity recognition framework based on wearable imu wrist sensors. In *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–5. IEEE, 2019.

[4] Pierluigi Casale, Oriol Pujol, and Petia Radeva. Human activity recognition from accelerometer data using a wearable device. In *Pattern Recognition and Image Analysis: 5th Iberian Conference, IbPRIA 2011, Las Palmas de Gran Canaria, Spain, June 8-10, 2011. Proceedings 5*, pages 289–296. Springer, 2011.

[5] Inc CiteDrive. Citedrive brings reference management to overleaf, 2022. https://www.kaggle.com/datasets/meetnagadia/human-action-recognition-har-dataset/data [Accessed: (Use the date of access)].

[6] Diane Cook, Kyle D Feuz, and Narayanan C Krishnan. Transfer learning for activity recognition: A survey. *Knowledge and information systems*, 36:537–556, 2013.

[7] AG Parth Goel and A AG. A survey on deep transfer learning for convolution neural networks. *Int. J. Adv. Sci. Technol*, 29(6):8399–8410, 2020.

[8] Fuqiang Gu, Mu-Huan Chung, Mark Chignell, Shahrokh Valaee, Baoding Zhou, and Xue Liu. A survey on deep learning for human activity recognition. *ACM Computing Surveys (CSUR)*, 54(8):1–34, 2021.

[9] Fuqiang Gu, Kourosh Khoshelham, and Shahrokh Valaee. Locomotion activity recognition: A deep learning approach. In *2017 IEEE 28th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)*, pages 1–5. IEEE, 2017.

[10] Fabio Hernández, Luis F Suárez, Javier Villamizar, and Miguel Altuve. Human activity recognition on smartphones using a bidirectional lstm network. In *2019 XXII symposium on image, signal processing and artificial vision (STSIVA)*, pages 1–5. IEEE, 2019.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[12] Oscar D Lara and Miguel A Labrador. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, 15(3):1192–1209, 2012.

[13] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[14] Sakorn Mekruksavanich and Anuchit Jitpattanakul. Exercise activity recognition with surface electromyography sensor using machine learning approach. In *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, pages 75–78. IEEE, 2020.

[15] Ling Pei, Songpengcheng Xia, Lei Chu, Fanyi Xiao, Qi Wu, Wenxian Yu, and Robert Qiu. Mars: Mixed virtual and real wearable sensors for human activity recognition with multidomain deep learning model. *IEEE Internet of Things Journal*, 8(11):9383–9396, 2021.

[16] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 270–279. Springer, 2018.

[17] Mohib Ullah, Habib Ullah, Sultan Daud Khan, and Faouzi Alaya Cheikh. Stacked lstm network for human activity recognition using smartphone data. In *2019 8th European workshop on visual information processing (EUVIP)*, pages 175–180. IEEE, 2019.

[18] Baoding Zhou, Jun Yang, and Qingquan Li. Smartphone-based activity recognition for indoor localization using a convolutional neural network. *Sensors*, 19(3), 2019.

[19] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.