National College of Ireland

Document Image Classification using Convolutional Neural Network and Transfer Learning Technique

MSc Research Project
Data Analytics

Asim Arif Ibushe
Student ID: x21172218

School of Computing
National College of Ireland

Supervisor: Mr. Bharat Agarwal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Asim Arif Ibushe
Student ID:	x21172218
Programme:	Data Analytics
Year:	2023-2024
Module:	MSc Research Project
Supervisor:	Mr. Bharat Agarwal
Submission Due Date:	25/04/2024
Project Title:	Document Image Classification using Convolutional Neural Network and Transfer Learning Technique
Word Count:	6535
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	25th April 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Document Image Classification using Convolutional Neural Network and Transfer Learning Technique

Asim Arif Ibushe
x21172218

Abstract

In todays digital age, efficent document classification and pattern recognition are essential for server-side big data structuring. This requirement is especially important for domain such as banking and government, where document processing and classification is essential. This study suggest an efficient and automated approach using deep learning and transfer learning technology to predict target label for document pictures and helps to automate image classification task. We have performed comparative study on two document image dataset to demonstrate which learning method suits the best. First dataset contain 6 different document image target label to classify, we have implemented CNN algorithm which performed 10% accuracy hike compared to traditional machine learning. Second dataset titled "Tobacco-3482" serve as a valuable resource containing total 3492 document images which demonstrates different patterns and structure including 10 separate directory for each target label. Research determines the incremental increase in prediction accuracy from traditional learning such as K-Nearest Neighbour, Support Vector Classifier to Convolutional and transfer learning technique. Inceptionv3 and VGG-16 algorithms results were compared with the help of Imagenet dataset weights. Apart from that, proposed system add a feature of data-deduplication python script, which keeps on observing all document records integrity and efficent directory storage optimization. According to post-analysis and accuracy score Inceptionv3 and VGG-16 are top performer with 40 epochs. Validation accuracy for Inceptionv3 and VGG-16 are 79.60% 81.61% respectively, which shows that Visual Geometry Group-16 performs the best for dataset "Tobacco3482".
Keywords: Image Convolution, Image Classification, Transfer Learning, Data Deduplication,KNN, SVM, Convolutional Neural Network, InceptionV3, VGG-16

1 Introduction

The need for digital data and its storage has grown critical for organizations in the current digital world, which makes it challenging to classify digital document image, particularly in government organizations, banking, health and finance. Manual document image classification and processing takes a lot of time and it is prone to mistakes at server and data warehouse, so this research find innovative methods to speed up these manual processes and increase accuracy score. Businesses operating in vital industries and government, need to deal with difficulties associated with handling huge document images of multiple class. There is a clear need for inventive approaches, especially when it comes to automating processes like document structuring and finding structural similarity using pattern

recognition to classify document images. This lowers the need for human resources and saves time by streamlining crucial manual document image classification procedures while also increasing the accuracy of image data segregation. In modern times of increased digital communication, document processing efficiency is critical to smooth operations in substantial data sectors for document verification. Our research work boost another new module to include feature like periodic document image data deduplication in the data storage center, with the goal of automating the image classification procedure. Creating an intelligent system to extract pattern from structured corporate identity document is the main goal. According to Chen(2021) as cited below Chen (2021), The best option is chosen after a comparative study of several deep and transfer learning algorithm which suits the best according to the distribution of the dataset, assuring that the suggested method successfully implements the document classification procedure. Taking into account some limitation of the system, we have to admit the depth of our research. The accuracy of the suggested cross-platform model may be affected by differences in the quality of the input image documents, which is a prerequisite for deep learning classification efficient effectiveness. The paper is structured to explore the problem associated with processing documents manually, highlighting the need of automating these procedures. Apart from that file system automation techniques to efficiently manage data storage system using hashing techniques such as message digest (md5) algorithm. Research implements concepts of transfer learning models which uses inheritance and weights and network parameters from imagenet, along with vital information extraction from physical document images to classify them. An in-depth understanding of the developments achieved in the fields of automated image processing and power of deep learning to label a particular document.

2 Related Work

A wealth of research into many applications and approaches reveals that the area of Image classification has improved significantly. This thorough literature study looks at a few publications that have all contributed to the expanding field of optical image recognition and classification. Human eyes notice various patterns and styles in genuine paper documents. Proofreading must be done manually in order to verify the document's card number. A scanned digital document is a graphics file having a pixel grid structure. A computer system can locate, identify, and recognise characters on an image document, enabling for the automated validation of thousands of documents.

S. Q. Gao delivered his presentation, "A Research on Traditional Tangka Image Classification Based on Visual Features," at the 2023 4th International Conference on Computer Vision, Image, and Deep Learning (CVIDL) in Zhuhai, China.It deep dive the process of categorizing traditional Tangaka images based on visual features. It explains concepts of feature extraction and classification technique, emphasizing the importance of image structure in Traditional Tangka image identification . Gao (2023)

The paper "Research on Image Classification Method Based on DCNN," presented at the 2020 International Conference on Computer Engineering and Application (ICCEA) in Guangzhou, China, by C. Ma, S. Xu, X. Yi, L. Li, and C. Yu, introduces a method for image classification using Deep Convolutional Neural Networks (DCNN). The research

explores at how effectively DCNN results in tasks involving image labelling. It focuses on the successful application of DCNN for accurately recognising pictures based on content. Author concludes suggested approach exceeds previous techniques, demonstrating DCNN's supremacy in handling complex image data. Study reflects importance of network architecture and training processes in achieving high classification accuracy. Ma et al. (2020)

H. Li's paper, "An Overview of Remote Sensing Image Classification Methods with a Focus on Support Vector Machine," presented at the 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML) in Stanford, CA, USA, provides a comprehensive analysis of remote sensing image classification techniques, with a focus on Support Vector Machine (SVM) methods. This paper analyzes several techniques for object and pixel based remote sensing image categorization. A comparison of pixel based and object based was compared. Additionally, insights into the value of SVM in dealing with remote sensing data are projected. Li (2021)

Narayana Kamath, Cannannore Nidhi, Bukhari, Syed, and Dengel, Andreas conducted a 2018 study titled "Comparative Study between Traditional Machine Learning and Deep Learning Approaches for Text Classification." This study compared the performance of deep learning to machine learning for classifying text. It outlines the advantages and disadvantages for each technique, how it impacts reflecting dataset characteristics on model performance, and few recommendations for selecting the best suitable algorithm corresponding a particular demands of text classification . Kamath et al. (2018)

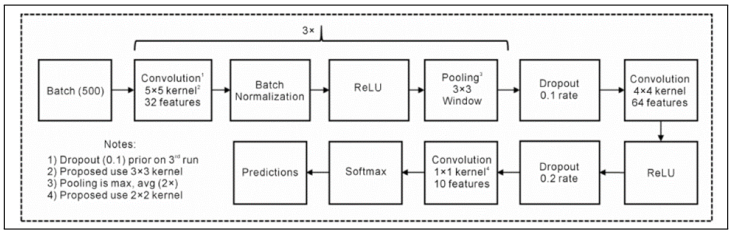

Figure 1: CNN structure for feature extraction
Source: `https://www.researchgate.net/publication/322957424_An_Ensemble_of_Convolutional_Neural_Networks_Using_Wavelets_for_Image_Classification`

Sharma and Phonsa's paper "Image Classification Using CNN" presented at the International Conference on Innovative Computing and Communication (ICICC) 2021, delves into the utilization of Convolutional Neural Networks (CNNs) for image classification tasks. The paper likely outlines the methodology, experimentation, and results achieved through CNNs in classifying images, demonstrating the efficacy of deep learning techniques in this domain. This research advances computer vision by proving how well CNN handle complicated graphic data and demonstrating their potential for a range of real-world applications, including autonomous driving and object detection and medical imaging. Additionally, the paper underscores the growing importance of deep learning approaches in image analysis, focusing their relevance in modern computational systems and artificial intelligence research. Sharma and Phonsa (2021)

Table 1: Comparison of Traditional Machine Learning and Deep Learning

Aspect	**Traditional ML**	**Deep Learning**
Model Complexity	Involves handcrafted feature engineering and selection	Automatically learns features from raw data
Performance on Large Datasets	Due to computing constraints, large-scale datasets may provide a challenge.	Excels at huge datasets by exploiting parallel computing power and hierarchical feature representations.
Generalization	If feature engineering is done properly, the results may be generalizable.	Can generalize effectively even with minimum feature engineering, especially when trained on a varied and representative dataset.
Interpretability	Frequently delivers interpretable models, offering insight into feature relevance and the decision-making process.	Models are sometimes seen as black boxes, making it difficult to grasp how decisions are made, especially in complicated structures.
Training Time	Training time may be shorter compared to deep learning approaches, especially for smaller datasets.	Due to the intricacy of neural network topologies, training requires substantial computer resources and time, especially for large-scale datasets.
Data Efficiency	It often takes less data to obtain good performance, making it suited for settings with minimal labeled data.	Frequently need enormous volumes of labeled data for training, which may not be viable in some applications.
Robustness to Noise	May struggle with noisy data, as feature engineering may amplify noise	May struggle with noisy data since feature engineering might enhance noise.

Williams and Li's paper, "Advanced Image Classification Using Wavelets and Convolutional Neural Networks," presented at the 2016 IEEE International Conference on Machine Learning and Applications (ICMLA) in Anaheim, CA, USA, looks into a hybrid approach for image classification tasks that combines wavelet transforms and Convolutional Neural Networks (CNN). This new technique certainly combines the remarkable feature of CNN with the location and frequency information capturing abilities of wavelet analysis. The paper's findings may highlight advances in classification accuracy and robustness achieved by this hybrid method, highlighting its potential to improve image analysis work in a variety of contexts. As a way to handle the challenges of image classification in real-world scenarios, this work highlights importing signal processing with deep learning. Williams and Li (2016)

Data deduplication issues in open clouds are addressed in following research. A method of 3 using many key servers and randomized tagging files in order to guarantee strong tag consistency is proposed by author Gosai and Das (2021).These help preventing the storing of identical documents data across cloud. Data deduplication technique optimize storage space economy while reducing these worries and boosting user trust in storing and retrieval of data. The proposed research invents prevention of fake file upload and guarantees robust tag consistency. Gosai and Das (2021)

The paper by Ju-Young Lee, Jae-Wan Lim, and Eun-Jin Koh, published in the Journal of Electrical Engineering and Technology, focuses on picture categorization using the HMC (Harmonic Mean Combination) approach combined with a Convolutional Neural Network (CNN) ensemble, particularly for infrared images. The effectiveness of this hybrid method in accurately recognizing infrared images, which are employed in a variety of corporate domains, including surveillance, medical imaging and environmental monitoring. The project aims to increase classification accuracy and robustness while dealing with complex infrared image data by combining the HMC technique with CNN Ensembles.This study highlights the necessity for utilizing cutting-edge strategies like CNN ensembles and fusion algorithms in order to enhance picture classification performance in particular domains like infrared imaging.. Lee et al. (2018)

The study by Song et al., published in IEEE Transactions on Image Processing, discusses a Deep Multi-Modal Convolutional Neural Network (CNN) for Multi-Instance Multi-Label Image Classification. The project is anticipated to look into the challenges of identifying numerous occurrences and assigning different labels to images at the same time, which are frequent in a variety of applications such as scene interpretation and medical diagnostics. The study proposes a multi-modal CNN architecture that can rapidly integrate data from several sources or modalities, hence improving classification accuracy and resilience. This study emphasises the need of developing CNN architectures for complicated picture classification problems, particularly in scenarios with many occurrences and labels, and so contributes to the larger area of computer vision and pattern recognition. Song et al. (2018)

Krishna, Neelima, Mane, and Matcha's paper, published in the International Journal of Engineering and Technology, focuses on photo classification using deep learning algorithms. The project is to examine the application of deep learning models, such as Convolutional Neural Networks (CNNs), for image classification tasks across several do-

mains. The study's purpose is to improve classification accuracy and efficiency over existing methods using deep learning. This work is significant because it adds to the progress of computer vision technology by suggesting potential solutions for tasks such as item recognition, medical imaging analysis, and autonomous driving systems, therefore addressing real-world problems in a wide range of disciplines. Manoj krishna et al. (2018)

Bukhari and Dengel's study, presented at the 13th International Conference on Document Analysis and Recognition (ICDAR) in 2015, looks on visual appearance-based document categorization techniques. The project will analyse and compare the performance of several methods for identifying documents based purely on their visual appearance, rather than their written content. This research is crucial because it adds to the development of robust document categorization systems, especially in cases when text extraction is difficult or impracticable, such as handwritten papers or documents in languages with complicated scripts. By comparing various methodologies, the study hopes to shed light on the benefits and limits of visual appearance-based approaches, which will inform future research in document analysis and identification. Bukhari and Dengel (2015)

Gupta, Ankit, Kulkarni, and Jain's work, published in 2022, looks at handwritten signature verification utilising transfer learning and data augmentation. The study is most likely investigating the use of employing pre-trained models and artificially created data to increase the accuracy and robustness of signature verification systems. This paper is significant because it addresses the challenges of properly validating handwritten signatures, which are required for authentication and security in a range of industries. The research aims to improve the performance of signature verification systems through transfer learning and data augmentation, hence contributing to advancements in biometric authentication technology. Gupta et al. (2022)

Pokharel and Giri's 2017 undergraduate project, "Offline Signature Verification Using Convolutional Neural Network," is likely an examination of the use of CNNs to authenticate offline handwritten signatures. The study may investigate the use of CNNs in evaluating signature images and determining their legitimacy without needing real-time involvement. This work is notable because it tackles the need for exact and reliable signature verification systems, which are essential in a range of areas including finance, law, and government. Using CNNs, the study seeks to contribute to the development of robust and efficient offline signature verification methods, hence increasing security and authentication protocols. Pokharel and Giri (2017)

Han, Liu, and Fan's study, published in 2018 in Expert Systems with Applications, introduces a novel image classification method employing CNN transfer learning and web data augmentation techniques. The research likely explores the utilization of pre-trained CNN models and additional data from the web to improve the performance of image classification systems. This work is significant as it addresses the challenges of limited labeled data and enhances the generalization capabilities of CNNs. By leveraging transfer learning and web data augmentation, the study aims to advance the accuracy and robustness of image classification tasks, offering potential solutions for various real-world applications in fields such as computer vision and pattern recognition. Han et al. (2018)

Wu,Liu, Long, and Hou's paper, presented at the 2019 IEEE 2nd International Con-

ference on Information and Computer Technologies, focuses on personal identification verification using a Convolutional Neural Network (CNN). The study will most likely investigate the feasibility and usefulness of CNNs for validating human identities, maybe using biometric data such as face photographs or fingerprints. This research is critical because it addresses the growing need for trustworthy and efficient identity verification systems, particularly in security-sensitive industries such as banking, border control, and access control. By studying CNN-based methodologies for personal identification verification, the study seeks to contribute to advancements in authentication technology, hence strengthening security measures across a wide range of applications.Wu et al. (2019)

3 Methodology

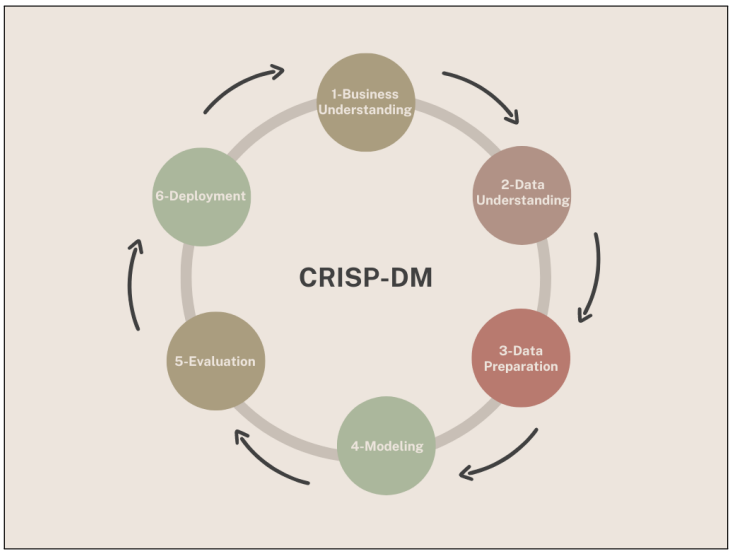

Figure 2: CRISP-DM Lifecycle

In my research project, I refereed CRISP-DM framework or (Cross-Industry Standard Process for Data Mining) which helped me to organize and guide me to generate building blocks. This technique helped me to simplify the process from project inception to result evaluation. It consist of six essential phases such as Business Understanding (Ask Domain experts), Data Understanding,also known as pre-analysis phase which include feature study and exploring subset of original dataset that gives an insights of patterns and structure of data. Next step is data preparation and dimensionality reduction also known as feature scaling and feature extraction. Data preparation phase also include dataset generalization and data split. To validate model in testing phase we have to preserve some test data from original dataset to calculate model performance. Hence train test split phase. Next phase in above CRISP-DM life-cycle is Modeling or algorithm implementation which include hyper-parameter tuning and architecture implementation. It is also known as model training phase. Trained model evaluation is the next phase where post-analysis is done. This involves calculating metrics and accuracy scores to evaluate the performance against the determined objectives. The final phase includes project deployment and regressive performance check. Integrating the Learning model in environment. This framework ensures that the project progress systematically.

3.1 Dataset Description and volume

Dataset Description: Dataset is the crucial part for a research study. As quality of data helps build ML model. In our case we have use two different dataset for document image classification. In first dataset named "Personal Identification"A variety of personal identity documents for Indian citizens are included in this public dataset generated by Mehak Singal. Numerous identity papers, including passports, PAN cards, Aadhaar cards are included in the collection. Total 600 to 700 personal identification document images are available for research purpose. Total target variable or labels for above directory are 6. As dataset is small in size we have considered it for post-analysis phase for comparative study of traditional and convolutional machine learning purpose. Singal (2023) The second dataset used is named "Tobacco3482" which is of size 1.8 gigabyte which contains total 3492 jpg image file format. Dataset is spread across 10 different directory each representing target variable (Labels). This dataset is complex and does not follow a routine pattern. We have used second dataset named "Tobacco3482" for comparitive study of user-defined CNN implementation using keras library and transfer learning technique such as InceptionV3 and VGG-16. Lewis et al. (2006)

3.2 Data Pre-processing and Data Augmentation

Image is represented as just a grid of pixels Ma et al. (2020). Hence each dot or bit is represented as just RGB colour notation from range of 0 to 255. The major task is to reduce image pixels size and binarization in grayscale image as a step of pre-processing and feature extraction Gao (2023). To Design a neural network for an input as image value contains a lot of input, for each of the pixel in an image we have numeric input from 0 representing white and 255 representing completely dark such as black in case of grayscale input.In case of first dataset we defined a function named as preprocess images() which reads images using OpenCv API, resize its to standard value 224*224 pixel count and binarize it by dividing each pixel value 255. While using transfer learning technique Inceptionv3 and VGG-16 we have defined same function which tranforms image to size of 244*244 and algo reads image in 3 channel format. After that to enumerate target field we have used LabelBinarizer() from scikit-learn to convert categorical target variable using one-hot encoded vector. It act same as fit transform and label encoder.

3.2.1 Data Augmentation for Large Dataset:

Numpy array was not suitable for dataset two as image count was huge and batch size had to be reduce due to memory limitation. Data augmentation is a method that applied different changed to input images in order to extract feature and make training data generalized. Pixel values are set up in the Image DataGenerator to be rescaled to a range of 0 to 1. Williams and Li (2016) We have defined object of ImageDataGenerator() class from keras library to apply binarization and rescale parameter for both training 80% data as well as testing 20%. Image size with 224 pixel matrix and batch size of 16.

Above two figure propagates Class distribution of both dataset, we can observe that dataset one has 6 distinct target labels and denotes neutral and fair count value. While in second dataset distribution we have plot it in ascending order which shows that target label count is skewed and more biased toward letter, mail and memo image document count.

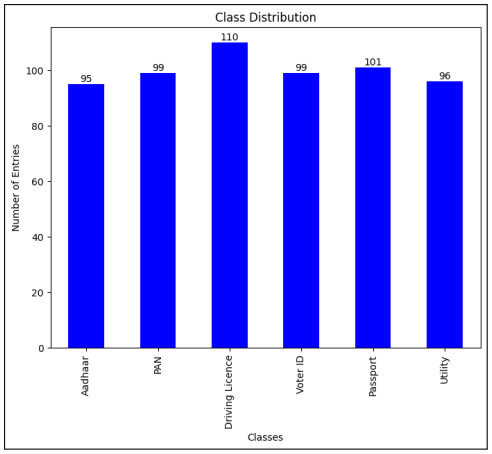

Figure 3: Class Distribution: Dataset 1

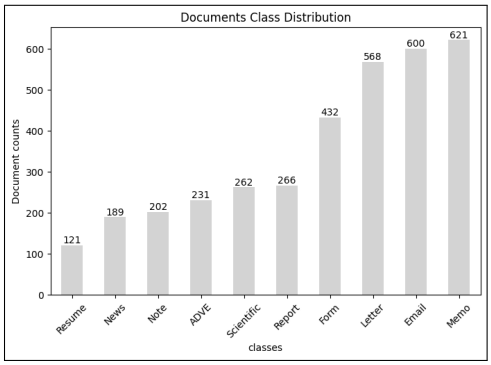

Figure 4: Class Distribution: Dataset 2

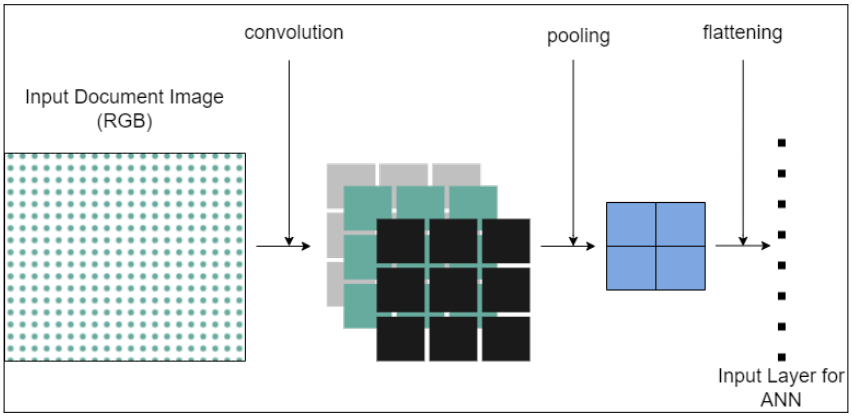

Figure 5: Convolution Intermediate Phase

3.2.2 Image Convolution:

Here to extract particular feature and curve in an image, Image convolution algorithm Designs helps for dimensionality reduction, its all about filtering an input image which help us to extracting useful and relevant features out of an image Lee et al. (2018). By applying a filter which outputs a pixels value based on its surrounding neighbors pixel values, reducing the size of an image. With the help of kernel or filter value already pre-initialized, filter is multiplied it with the window matrix of actual image to reduce it to single value. Hence we generate a feature map with the help of multiplying kernel value with different image region and stride jump. Hence if focused image pixel find deviating from its neighbouring image pixel value from its window kernel, finds helpful in detecting horizontal or vertical edges in an image.Image convolution helps in extracting certain useful results out of those image. Another technique know as pooling is used to sampling from different region of an image resulting in dimension reduction. Max-pooling can be used to select a maximum value from a particular region matrix, helps in shrinking a size of an image pixel grid. Hence by aggregating Image convolution, pooling and traditional neural network, image classification technology uses convolutional neural network algorithm usually in the context of analyzing and predicting an image document.Sharma and Phonsa (2021) Deep neural network represents 6 neuron in the output layer that tells us what it predicts 6 different multi-class target label and 10 unique

document structure in case of "Tobacci3482" second dataset document classification.

3.3 Modelling and Defining Network Architecture

3.3.1 K-Nearest Neighbour:

KNN is a lazy learning algorithm, it just plots all the traning data points and corresponding target labels to memory. KNN computes the distance between each new point need to predict label with every other point in the training set. K represents the hyper-parameter and it's suggested to define it as Odd integer value. New data point's predicted label is determined by taking the most frequent class label among its k neighbours using polling method. KNN uses euclidean distance as a metrics to calculate actual distance. Hence in our code n_neighbors optional parameter is set to 5. Selecting a suitable value for k is essential as it may greatly affect the model's performance and capacity for stability. Though it is categorical classification algorithm, it does not best suits for image classification as it does not recognizes pattern, curves and edges. Convolutional step is must in case of image datasets. Kamath et al. (2018)

3.3.2 Support Vector Machine:

In SVM classifier training data is represented as points in space that are divided into categories by large gap as possible. Categories are separated by a line known as hyperplane. In our code parameter kernel=linear uses a linear kernel to discover the best linear decision boundary between plane space dimension. The regularization parameter c, helps lowering classification error and maximizing the margin using supporting vector at the edges. Hence c determines the size of margin and acts as a hyper-parameter. SVM gains learning of the ideal decision boundary during training that best divides the various classes within the feature space memory. We have set value of C=1.0, which is optimal in nature. Hence, helps in reducing over-fitting problem. Li (2021)

3.3.3 Inception V3:

Inception Version3 is CNN architecture developed by Google. It is primarily designed for image classification task and uses a series of convolutional layers with different filter sizes such as (1*1,3*3,5*5) which are attached together in parallel. Pre-trained weights are available from InceptionV3 from large-scale dataset like ImageNet allowing for progressive transfer learning on concatenated custom user-defined network. With pre-trained weights and bias from ImageNet dataset, we have initialized the parent model with the InceptionV3 architecture. We have set include_top=Fasle as an argument which excludes top layer and allows extended model to add our own classification layers.

3.3.4 Visual Geometry Group (VGG-16):

VGG is a CNN architecture developed at the University of Oxford. It consists of 16 weight layers, including 13 convolutional layer and 3 fully connected layers.It is famous for its efficiecy and ease of use in image categorization applications. By reducing the number of dimensions and avoiding overfitting, it was designed using small size filters 3x3 size window and followed by max-pooling layers of size 2x2 improve computing efficiency. VGG-16 internal architecture include 3 fully connected layer followed by convolutional steps use rectified linear unit(ReLU) activation function, each layer with 4096 neurons.

The last output layer uses softmax activation, with 1000 neurons representing ImageNet classes. In transfer learning, the top-most fully connected layer is eliminated and replaced with a layer designed for the particular user-defined classification job,making use of the pre-trained weights and bias of VGG-16. Kamath et al. (2018)

3.4 Document Image Deduplication using MD5 hashing technique

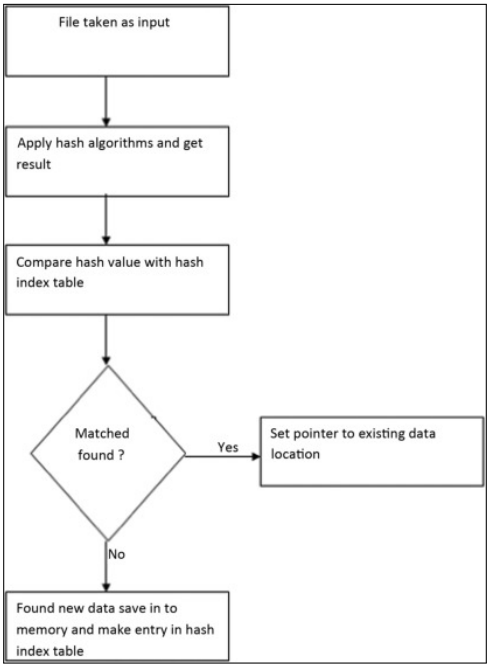

Figure 6: Absolute Path Dictionary Flow, Source:`https://www.sciencedirect.com/science/article/pii/B9780128233955000021`

Document images stored in two dataset directory are redundant in nature leads to large store consumption, there was a need for document integrity check and storage optimization technique. Redundant and duplicate image data file results in inefficient storage structuring and duplicate record in model training phase. Using the MD5 hashing technique create distinct hash values for every input image document file known as blob. The 128-bit MD5 has is commonly expressed as a 32-character hexadecimal number. We calculated checksum value of each image file using bucket data block extraction technique and stored checksum value as key in dictionary. File absolute path will act as value parameter to each checksum key in dictionary structure. System periodically iterates through dictionary initialized file-path value and if filepath value count is greater than 1 for any checksum key, it's automatically dropped and removed by operating system automation code, the duplicate document image with same file cheksum value will be deleted maintaining storage integrity and efficiency Gosai and Das (2021).

4 Design Specification

Computer method for analyzing and understanding digital images is computer vision Gosai and Das (2021). Hence, classification techniques is applied to distinguish differ-

ent document images and pattern recognition using self feature extraction techniques of neural network. Wu et al. (2019) Internal Back-propagation techniques is used by CNN to tune and set hyperparameters such as weights and bias between intermediate nodes in fully connected layer

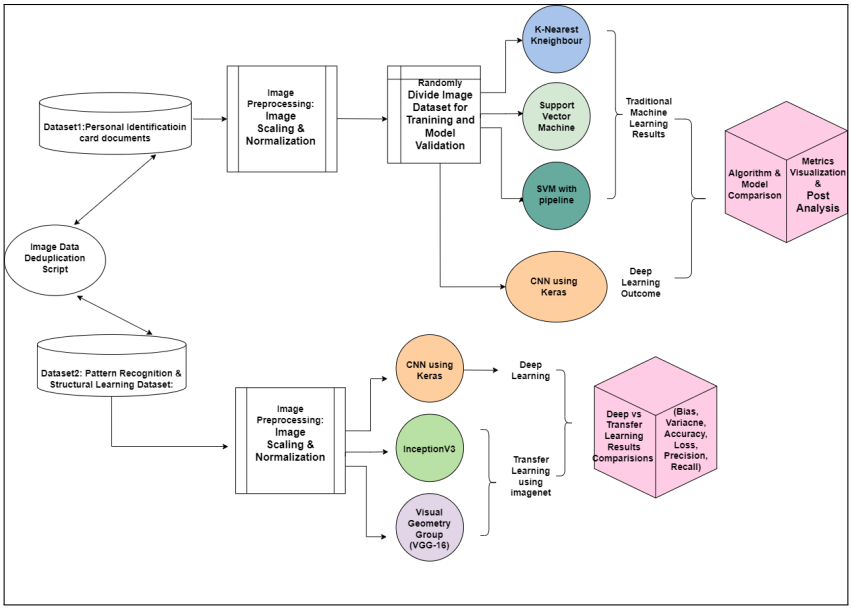

Figure 7: Implementation Flow Chart

Below Fig:7 Flow chart demonstrate comparative study for many learning algorithm on two different dataset. Image data deduptication python script is applied for two different dataset directory. Further Image pre-processing and image scaling technique are used for normalization before training phase. K-nearest neighbour, Support Vector Machine, Keras CNN algorithm were experimented on personal identification dataset 1. Results were compared and parameter such as value of k, kernel value in SVM are tunned according to best results. For dataset2: "Tobacco-3482", we designed keras CNN with 10 epoch value count. But results were not satisfying. We earlier kept batch value of 32. But due to memory limitation we reduced its value to 16. Dataset 2 was scattered and was not structural in nature so we had to implement transfer learning technique. we have inherited and utilized transfer learning. The intentional transfer of already trained parameter from known dataset to another is known as transfer learning. In simple terms, vital information such as weights and network architectural parameters are transmitted and used to train future models once the parent dataset such as (Imagenet) has been trained on the existing model such as InceptionV3 and VGG-16 as in above figure. It was evident from detailed analysis of literature that applying this approach also resulted in dramatic improvements for poor image dataset in model accuracy. All model result are compared for deep understanding and to conclude which suits best for huge image data classification.

Top most layer of pre-trained model are dropped to append and extend user-defined fully connected layer dence layer followed by convolutional architecture. To overcome the over-fitting problem dropout layer is added and softmax activation function is applied to output layer for categorical classification of target document image label such as email, letter, reports, news.

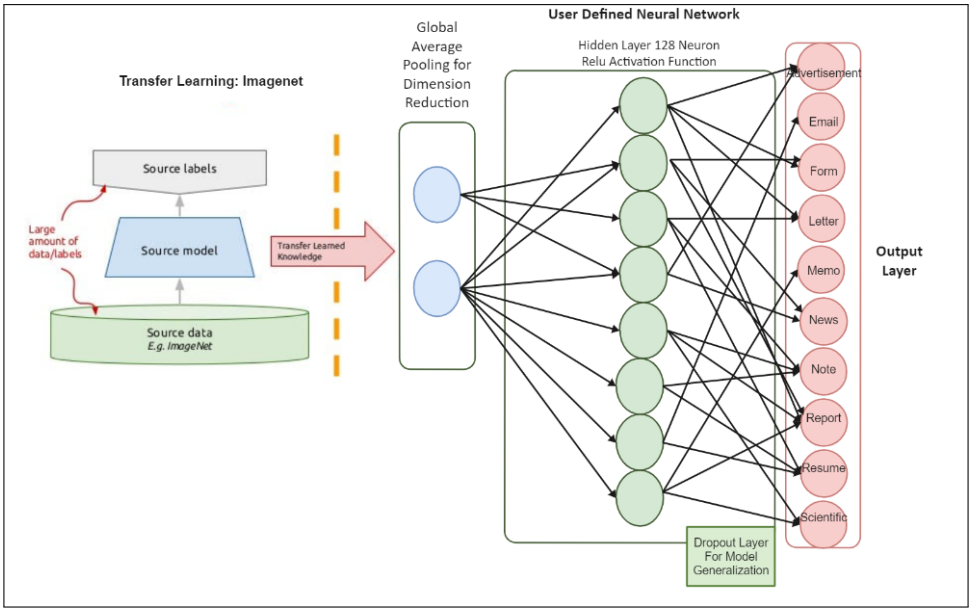

Figure 8: Transfer Learning Architecture

5 Implementation

After pre-processing step elements are removed from document images and grayscale and binary threshold is achieved Gao (2023).

In Data preparation step we did random splitting of image dataset along with its label value at random state of 42 unit, test size is 20% for model validation and calculating accuracy score for cnn implementation on Dataset 1.

Building CNN model and training phase: Using keras library sequential class instant creation Manoj krishna et al. (2018). Training CNN with 32 unique filters and relu activation function. We have transformed each document image into 224*224 dimension and 3 channel representing RGB value in coloured image.

Compiling the model with optimizer adam which is responsible for efficient updation of weights during training phase to minimize loss rate. categorical cross-entropy is used as loss parameter for multi-class classification.The neural network model is configured for training in the compile stage.

Model building phase: we have set epochs value as 8, epoch determines the number of times the entire training dataset is back-propagated and weight adjustment is done to minimize the loss rate Lee et al. (2018). We can notice that Accuracy scored got increase from 18.54% from 1 epoch round to 84.17% in the last epoch phase. We can save trained model to preserve its parameter and trained network structure. we saved our model in a file format named "image_classifier_model.h5" As I tried to increase epoch count, training accuracy kept to be constant but validation accuracy dropped resulting in high variance. Model was pretending to be over-fitted if epoch count was increased more than 8 for dataset1.Singal (2023)

Trained CNN Model Dataset 1: There are total three convolutional layer, following every convolutional layer is max-pooling layer. First filter size is 32, second is 64 and last third size count is 128 respectively. Flattening layer is used to transform output to a single dimension vector. For document image categorization last dense layer are set to count 6 or len(class labels), softmax activation function is applied to output layer for

multiclass categorical label predictions.

Total training time was approximately 2 minutes and 8 seconds. There were total 8 epoch round, each epoch denotes a single entire run through training set back and forth.

As number of epoch count keeps increasing, more number of feature are extracted and helps for pattern recognition. We can observe that Train as well as test accuracy keeps on increasing, resulting in low bias and low variance. Effects in generalized deep model without any over or under fitting problem.

InceptionV3 Implementation: As dataset2 is not accurate and limited training data available, we have implemented transfer learning using Inceptionv3 pre-trained weights as our base model. We have instantiated our base model with the help of InceptionV3 constructor assigning pre-trained weights from ImageNet dataset. We have set include_top parameter to false indicating that we will be integrating our own custom categorization layers. input_shape accepts tuple defining image size as 224*224 and channels value as 3 indicating RGB. Parent model InceptionV3 generates sequence of feature map along with pre-trained weights and bias value. GlobalAveragePooling2D() method computes the average of each feature map reducing overall image dimensions still maintaining integrity. Next we have added hidden layer with 128 neuron count with relu activation function resulting non-linearity in intermediate nodes. To avoid overfitting problem we have added dropout layer with value 0.5, randomly eliminating decision neurons during training. To design output layer Dense() method is used with first parameter matching the number of classes in image_label target value from our dataset. Second parameter as Softmax activation function for multi-class classification. Next step is compiling the model where we have used Adam optimizer with learning rate of 0.00001 and loss set to categorical crossentropy for multi-class classification. In training phase fit() method train model using optional parameter training data and validation data with epoch value as 40 and batch size as 32. Validation data is passed to calculate Bias and variance parameter to analyse model generalization. Han et al. (2018)

VGG-16 Implementation: We have used VGG-16 transfer learning technique on "Tobacco-3482" dataset2. Same as above used pre-trained weights from imagenet dataset and set last output layer as False, so we can append custom network model further. While creating custom model we have initialized sequential model where I can stack additional layers. Further in architecture we have implemented hidden layer with 128 neurons with relu activation and added a droupout with .5 paramter as input to reduce overfitting. Last Adding last output layer with len of target variable size in our case it is 10 with softmax activation function. Ye et al. (2021)

Image Data Deduplication Algorithm Implementation: For warehouse deduplication of document images, design implementation is divided into three sub-modules. calculateChecksum(), DirectoryTraversal() and DeleteDuplicate()

First step is to read the file in binary using bucket blocks of 1024 byte size. compute the md5(message digest) hash code and return back hexadecimal checksum to get preserved. So keeping the track of duplicate files and there corresponding checksum value using key, value pair using directorytraversal sub-module. DeleteDuplicate uses system automation and ask admin consent to delete duplicate files. Below are the output prompt which showcase terminal output where total hash calculated for dataset 1 are 606, from which 36 Document Images were deleted from 6 different directories. Same for 3492 documents images checksum() value are calculated and total duplicate file count found was only 5. Python script ask user-prompt for Yes or NO to delete these duplicate images.

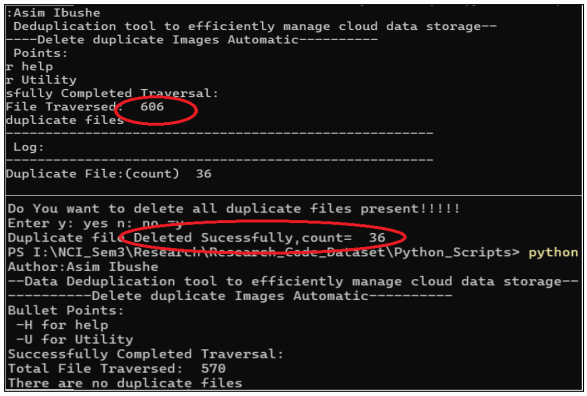

Figure 9: Deduplication Dataset1 Log

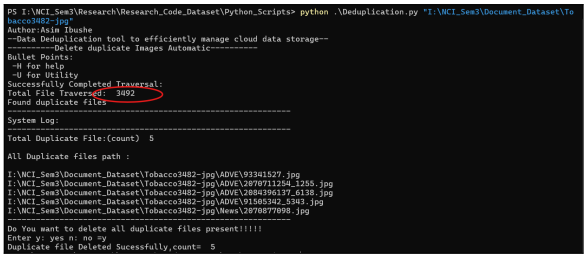

Figure 10: Deduplication Dataset2 Log

6 Evaluation

6.1 Learning Result for Dataset 1:

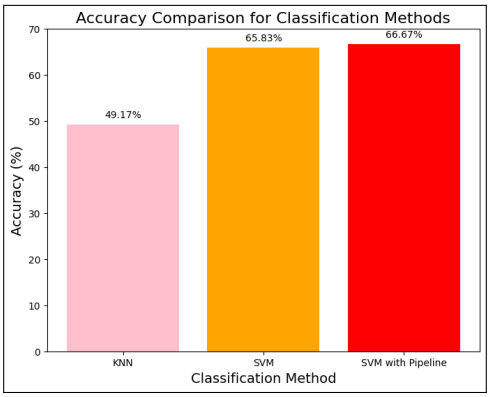

Figure 11: Dataset 1 Traditional ML: Accuracy Compare

Above Bar graph showcase that after self feature extraction and applying traditional ML algorithm result are poor on image dataset. K-nearest neighbour performed least with accuracy score of 49.17%. SVM with pipeline shows slightly high accuracy score as compared to normal SVM with linear kernel value. Using standardscaler() instant creation pipeline helps in preprocssing image in vector for better results. Output shows that we need to apply convolutional neural network for image classification.

Below two line graph shows accuracy and loss results of CNN for dataset1. Results in accuracy hike of 10% in validation phase. Loss visualization results that it dropped down from 1.73 to 0.73 till the last epoch count.

Confusion matrix interprets that actual true and predicted true are high in count, all diagonal True positive value are notably high. Overall results that F1 score is high. Error rate is seen to be low. We have calculated Bias and variance, which were 13.7 and 10.49 respectively. Shows that both are average. Model seems to better that previous traditional ML algorithms

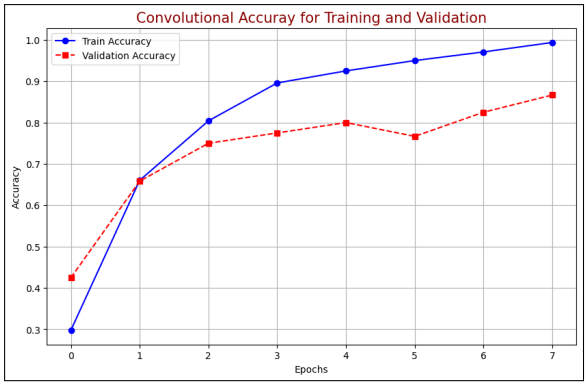

Figure 12: Accuracy Score

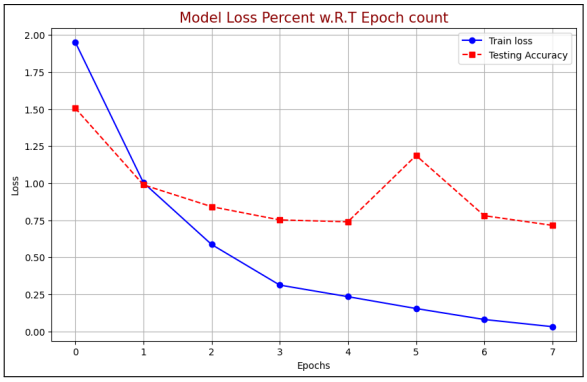

Figure 13: Loss Rate

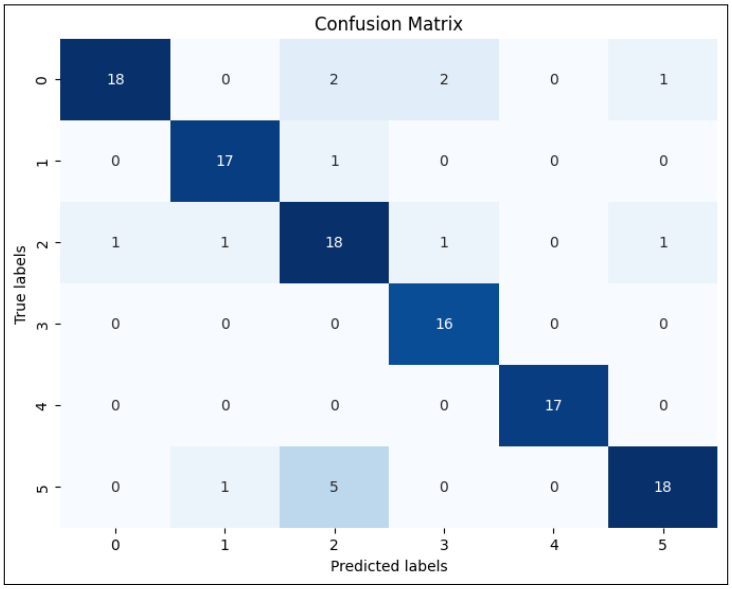

Figure 14: Dataset 1: Confusion Matrix for CNN Implementation

6.2 Learning Result and Post-Analysis for Dataset 2:

We have implemented CNN using keras for dataset 2, as dataset size is huge and total 3400 images. CNN did not performed well. After completing total 10 epoch with batch size of 16, we could only achieve highest validation accuracy of 56.98%. Instead validation loss kept on increasing from 1.40 in first epoch to 2.88 in last epoch. Line graph denotes its a overfitted model where training accuracy reached 98.71% and testing accuracy kept steady at range of 50%. Hence, denotes Low bias and High variance. Study suggest that there was a need to extend this experiment with concept of transfer learning.

After applying 40 epochs using InceptionV3 pretrained model and inheriting weights from imagenet dataset, accuracy for training as well as validation seems to improving. As dataset 2 is scattered and not regularized, still model is learning from training data of base in an efficient manner. Adam optimizer helps adjusting weights there is no signal of over-fitting model. Model boosts Training accuracy and minimizes over-fitting at the same time.Epochs 1 to 15 exhibiting a dramatic improvement, a decrease in learning rate was observed after 20 epochs and final accuracy observed for InceptionV3 model was 79.60%

Both InceptionV3 and VGG-16 somewhere result same accuracy output at 40 epoch

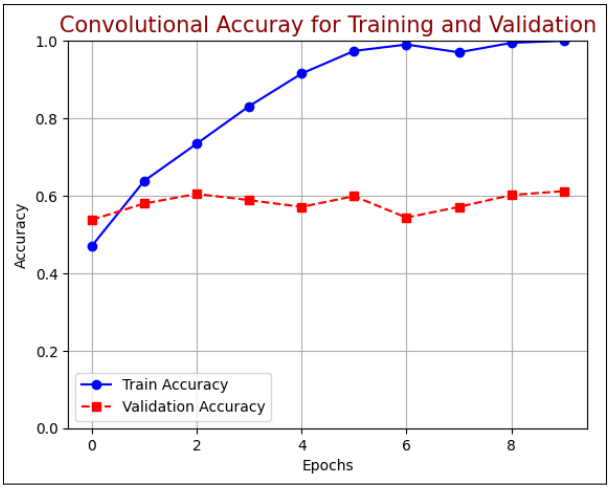

Figure 15: Dataset 2: CNN using Keras

```
Epoch 35/40
2437/2437 [==============================] - 28s 11ms/step - loss: 0.0352 - acc: 0.9938 - val_loss: 0.8213 - val_acc: 0.7989
Epoch 36/40
2437/2437 [==============================] - 28s 11ms/step - loss: 0.0363 - acc: 0.9930 - val_loss: 0.8243 - val_acc: 0.8017
Epoch 37/40
2437/2437 [==============================] - 28s 11ms/step - loss: 0.0330 - acc: 0.9951 - val_loss: 0.8394 - val_acc: 0.8046
Epoch 38/40
2437/2437 [==============================] - 28s 11ms/step - loss: 0.0257 - acc: 0.9967 - val_loss: 0.8449 - val_acc: 0.7859
Epoch 39/40
2437/2437 [==============================] - 28s 11ms/step - loss: 0.0244 - acc: 0.9963 - val_loss: 0.8497 - val_acc: 0.8046
Epoch 40/40
2437/2437 [==============================] - 28s 11ms/step - loss: 0.0282 - acc: 0.9959 - val_loss: 0.8525 - val_acc: 0.7960
```

Figure 16: Dataset 2: Accuracy and Validation Results InceptionV3

value. Accuracy Hike was seen to be tremendous in first 10 epoch count. Later it remained constant in range of 75%. VGG-16 seems to slightly dominant as compare to InceptionV3. Both model used Imagenet dataset base model weights for tranfer learning for image dataset. Final accuracy result achieved for Training accuracy was 80 to 81 percentage.

```
Epoch 35/100
2437/2437 [==============================] - 20s 8ms/step - loss: 0.0625 - acc: 0.9803 - val_loss: 0.8700 - val_acc: 0.8075
Epoch 36/100
2437/2437 [==============================] - 20s 8ms/step - loss: 0.0538 - acc: 0.9795 - val_loss: 0.9815 - val_acc: 0.8017
Epoch 37/100
2437/2437 [==============================] - 19s 8ms/step - loss: 0.0484 - acc: 0.9860 - val_loss: 0.9555 - val_acc: 0.8060
Epoch 38/100
2437/2437 [==============================] - 19s 8ms/step - loss: 0.0422 - acc: 0.9856 - val_loss: 0.9891 - val_acc: 0.8190
Epoch 39/100
2437/2437 [==============================] - 19s 8ms/step - loss: 0.0381 - acc: 0.9881 - val_loss: 1.1335 - val_acc: 0.7974
Epoch 40/100
2437/2437 [==============================] - 20s 8ms/step - loss: 0.0603 - acc: 0.9791 - val_loss: 0.8880 - val_acc: 0.7945
```

Figure 17: Dataset 2: Accuracy and Validation Results VGG-16

Inceptionv3 has more complex internal architecture and requires more computational power. InceptionV3 is generally opted for object detection and known for capturing specific details in image. While VGG16 is has been known for task such as image categorization. Both these tranfer learning techniques overcome the problem of overfitting and helps make trained model generalized in application. We tried to changed epoch value for both transfer learning technique but accuracy remained constant at certain level. Above table demonstrate that after data augmentation and jumping from normal CNN implementation to transfer learning techniques, got an improved results in accuracy and less gap between training error and validation error. For large and complicated dataset CNN lacks is ability, whereas transfer learning boost performance accuracy somewhere

Evaluation	CNN	Inceptioinv3	VGG-16
Training Accuracy	89.33	99.40	97.90
Validation Accuracy	61.70	79.60	81.61
Epoch Count	10	40	40
Data Augmentation	YES	YES	YES

Figure 18: Dataset 2: User-defined Deep vs Transfer Learning Results

reaching near 80 percentage.

7 Conclusion and Future Work

In conclusion, the study offers an in-depth approach for comparative study for traditional machine learning algorithms such as KNN, SVM with Convolutional deep learning on two different document image datasets. Research address the lack of expertise in transfer learning or inherited learning using InceptionV3 and VGG-16 for document image classification and provide feasible options for a wide variety of corporate sectors for automating digital document image classification. Research aimed to assess the performance of two ML model and three deep learning model, presenting benchmark for further research and practical applications in corporate business. Additionally, the system uses MD5 hashing for document deduplication to handle data redundancy in directory structure. System exhibits a modularized and organized structure defining is quality for reusability. Result demonstrates how effectively the model generalizes and reduces prediction errors. The quality of dataset is crucial in the domain of document image categorization, better preprocessing techniques for images helps handling pattern recognition and structural classification. Image superiority affects the overall performance of the system. We discovered that Both InceptionV3 and VGG-16 were top performing and traditional ML definitely does not suit image dataset. Even with the positive metrics, there continues to be room for advancement. In future work, more transfer learning algorithms can be implemented and we can integrate this technology with server-side automation script in many business. Finally, collecting more wide range of similar document images with better structural pattern may improve the learning ability to generalize output for document classification.

8 Acknowledgement

This study gratefully acknowledges the guidance given by the project supervisor Mr. Bharat Agarwal. His continuous guidance and support have contributed notably in this work and have led to significant improvements throught my research work.

References

Bukhari, S. S. and Dengel, A. (2015). Visual appearance based document classification methods: Performance evaluation and benchmarking, *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, pp. 981–985.

Chen, T. H. (2021). *An Artificial Intelligence Based Approach to Automate Document Processing in Business Area (Doctoral dissertation)*.

Gao, S. Q. (2023). A research on traditional tangka image classification based on visual features, *2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, IEEE, pp. 13–16.

Gosai, B. and Das, M. L. (2021). Data deduplication scheme with multiple key servers in public clouds, *2021 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, IEEE, pp. 272–277.

Gupta, Y., Ankit, Kulkarni, S. and Jain, P. (2022). Handwritten signature verification using transfer learning and data augmentation, *Proceedings of International Conference on Intelligent Cyber-Physical Systems: ICPS 2021*, Springer, pp. 233–245.

Han, D., Liu, Q. and Fan, W. (2018). A new image classification method using cnn transfer learning and web data augmentation, *Expert Systems with Applications* **95**: 43–56.

Kamath, C. N., Bukhari, S. S. and Dengel, A. (2018). Comparative study between traditional machine learning and deep learning approaches for text classification, *Proceedings of the ACM Symposium on Document Engineering 2018*, pp. 1–11.

Lee, J.-Y., Lim, J.-W. and Koh, E.-J. (2018). A study of image classification using hmc method applying cnn ensemble in the infrared image, *Journal of Electrical Engineering and Technology* **13**(3): 1377–1382.

Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D. and Heard, J. (2006). Building a test collection for complex document information processing, *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 665–666.

Li, H. (2021). An overview on remote sensing image classification methods with a focus on support vector machine, *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*, IEEE, pp. 50–56.

Ma, C., Xu, S., Yi, X., Li, L. and Yu, C. (2020). Research on image classification method based on dcnn, *2020 International Conference on Computer Engineering and Application (ICCEA)*, IEEE, pp. 873–876.

Manoj krishna, M., Neelima, M., Harshali, M. and Venu Gopala Rao, M. (2018). Image classification using deep learning, *Int. J. Eng. Technol.* **7**(2.7): 614.

Pokharel, S. and Giri, S. (2017). Offline signature verification using convolutional neural network (undergraduate project [subject code: Ct 707]).

Sharma, A. and Phonsa, G. (2021). Image classification using CNN, *SSRN Electron. J.* .

Singal, M. (2023). Personal identification image dataset for india, https://www.kaggle.com/datasets/mehaksingal/personal-identification-image-dataset-for-india. Accessed on 13 November 2023.

Song, L., Liu, J., Qian, B., Sun, M., Yang, K., Sun, M. and Abbas, S. (2018). A deep multi-modal cnn for multi-instance multi-label image classification, *IEEE Transactions on Image Processing* **27**(12): 6025–6038.

Williams, T. and Li, R. (2016). Advanced image classification using wavelets and convolutional neural networks, *2016 15th IEEE international conference on machine learning and applications (ICMLA)*, IEEE, pp. 233–239.

Wu, J., Liu, C., Long, Q. and Hou, W. (2019). Research on personal identity verification based on convolutional neural network, *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)*, IEEE, pp. 57–64.

Ye, M., Ruiwen, N., Chang, Z., He, G., Tianli, H., Shijun, L., Yu, S., Tong, Z. and Ying, G. (2021). A lightweight model of vgg-16 for remote sensing image classification, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**: 6916–6922.