Computing, Digital Forensics and Cybersecurity

BSc (Hons.) Computer Science

2020-2021

IS 20

Title: *Analysing the Correlation between Convolutional Filter Sizes and Classification Accuracy Using MNIST*

Author: *Gentian Gashi*

Supervisor: *Dr Richard Henson*

E-mail: *g.gashi193@canterbury.ac.uk*

This report is submitted in partial fulfilment of the requirement for the BSc in Computer Science at Canterbury Christ Church University

I declare that this report is my own original work containing no personal data as defined in the Data Protection Act (1998) and that I have read, understood and accept the University's regulations on plagiarism/intellectual property rights/research ethics (in particular the Research Governance Handbook) and the IS 20 Module Handbook.

Further, I accept that digital and/or hard copies of my Individual Study 20, or parts thereof, may be made available to other students, individuals and organisations after it has been marked.

Finally, I accept that no copy of my Individual Study 20 will ever be returned regardless of the circumstances.

Signed:

Date of Submission: 14th May 2021

Abstract

Handwriting recognition is the process of automatically converting handwritten text into electronic text (letter codes) usable by a computer. The increase in technology reliance during an international pandemic caused by COVID-19 has showcased the importance of ensuring the information stored and digitised is done accurately and efficiently. Interpreting handwriting remains complex for both humans and computers due to the various styles and skewed characters. In this study, we conducted a correlational analysis on the association between filter sizes and the convolutional neural networks (CNN's) classification accuracy. The testing has been conducted from the publicly available MNIST database of handwritten digits (LeCun and Cortes, 2010). The dataset consists of a training set ($N = 60,000$) and a testing set ($N = 10,000$). Using ANOVA, our results indicate a strong correlation ($.000, P \leq 0.05$) between filter size and classification accuracy. However, this significance is only present when increasing the filter size from 1x1 to 2x2. Larger filter sizes were insignificant therefore, a filter size above 2x2 cannot be recommended.

Keywords: *Handwritten Digits Recognition, Convolutional Neural Network, Convolutional Filter Size, Deep Learning*

Acknowledgements

I would like to thank my supervisor, Dr Richard Henson and my tutor, Dr Vijay Sahota for their invaluable advice, continuous support, and patience during my BSc study. My gratitude extends to all staff in the department of computing, this study would not have been possible without their continued help and support throughout all phases of this project.

Table of Contents

Abstract.....	i
Acknowledgements.....	ii
List of Figures	v
List of Tables	vi
1. Introduction	1
1.1. Introduction	1
1.2. Aims and Objectives	2
2. Literature Review	2
2.1. Introduction	2
2.1. A Brief History of Pattern Recognition.....	3
2.2. CNN Architectures	4
2.3. Conclusion.....	4
3. Premise	5
4. Methodology.....	7
4.1. Environment Setup.....	7
4.2. MNIST Dataset	7
4.3. Network Architecture.....	8
4.4. Pre-Processing	8
4.5. Training and Testing Data.....	9
5. Findings	10
5.1. Experiment Results	10
5.2. Estimated Marginal Means.....	11
5.3. Post Hoc Tests.....	13
5.4. Tests for Normality	14
6. Evaluation.....	15
7. Discussion.....	15
8. Legal Considerations	16
8.1. Health and Safety at Work Act 1974	16
8.2. Computer Misuse Act 1990.....	17
8.3. Data Protection Act 1998/2018.....	17
9. Ethical Considerations.....	17
9.1. Ethics Review Checklist.....	17
9.2. Unethical Future Research	17

10. Conclusion	18
References.....	19
Bibliography	22
Appendix A: Glossary	I

List of Figures

Figure 1: Rosenblatt's Simple Perceptron Model (Rosenblatt, 1960)	3
Figure 2: Identity Shortcut Connection (He et al. 2016).	4
Figure 3: Fishbone Diagram	5
Figure 4: 28x28 Pixel Digit from MNIST	7
Figure 5: CNN Architecture	8
Figure 6: Mean Accuracy by Filter Size Line Chart.....	12

List of Tables

Table 1: Software Versions Used.....	7
Table 2: Distribution of Digits in MNIST Dataset	7
Table 3: Descriptive Statistics on Filter Sizes	10
Table 4: Test of Between-Subject Effects.....	11
Table 5: Estimated Mean of Filter Sizes	11
Table 6: Multiple Filter Size Comparison.....	13
Table 7: Homogeneous Subsets	14
Table 8: Normality Test	14

1. Introduction

1.1. Introduction

Artificial intelligence (AI) resembles a renaissance after several false starts. Increased computational resources, explosive growth in data, a focus on specific problems, knowledge engineering with AI and alternative reasoning models, have all contributed to the emergence of intelligent machines (Hammond, 2015).

The idea surrounding automated handwriting recognition emerged in the 1950s due to research interest in reducing the substantial cost of getting information digitised (Dimond, 1958) as well as interest in recognising disguised and forged handwriting (Leedham, C.G., 1994).

There are two types of classes in handwriting recognition, online and offline. Online handwriting recognition systems work in real-time, processing and classifying handwriting styles as they are being written. Offline handwriting recognition systems process and classify characters only after it has been written (Leedham, C.G., 1994). Despite the increased popularity in handwriting recognition, it still faces challenges such as the various shapes and sizes of the characters as well slanted text found in cursive writing (Michal, Kozielski, Doetsch and Ney, 2013).

Handwriting recognition has become vital as it allows commercial sectors such as banks, to take advantage of automatic processing systems to process cheques, improving scalability, cost, automation, and accuracy by minimizing human error, as well the ability to allocate resources and manpower onto more complex problems. Other applications such as recognising forged handwriting could aid and improve security against identity fraud within organisations, minimizing human error by recognising patterns and features that may have been missed (Leedham, C.G., 1994). The emergence of the COVID-19 pandemic has forced organisations to adapt using remote working, further showcasing the need and importance of improving the quality, productivity, and security of organisations at a minimum cost.

Recent advances in neural network techniques have allowed researchers to create models such as long short-term memory (LSTM), which is a type of recurrent neural network (RNN) and connectionist temporal classification (CTC), which is a type of neural network output used to train RNNs, resulting in the algorithm being able to recognise handwriting styles despite the very limited and noisy datasets used (Pham et al., 2020).

1.2. Aims and Objectives

The main objective of this study is to compare various filter sizes to explore whether filter size has any significance on the model's classification accuracy. The results will be used to make a recommendation on which filter size should be considered for handwritten digits recognition.

Research Objectives:

- (RO1) Evaluate the accuracy of convolutional filter sizes for the development of handwritten digits recognition.
- (RO2) Recommend a filter size for handwritten digits recognition.

Research Questions:

- (RQ1) What is the relationship between filter sizes and classification accuracy?
- (RQ2) Which filter size should be recommended for the development of handwritten digit recognition using the MNIST dataset?

2. Literature Review

2.1. Introduction

Internet usage has almost doubled in 2020 due to the international pandemic, with Openreach reporting 50,000 petabytes of data consumed, compared to the 22,000 petabytes in 2019 (Openreach, 2021). This increasing reliance on technology has increased the importance and need for information being stored and passed digitally to be as accurate and efficient as possible. The following review of literature is intended to go through the history of pattern recognition and highlight different methods of handwritten digit recognition proposed throughout history. This will be used to build a foundation for a methodology to be used later in the study. The direction of this research could lead to the optimisation of document recognition and real-time OCR techniques. The optimisation of these areas could lead to the use of transferring sections of documentation online. The first section in the literature review 'A Brief History of Pattern Recognition' will highlight historic advancements in pattern recognition and the algorithms used. The second section 'CNN Architectures' will cover variations and advancements in architectures.

2.1. A Brief History of Pattern Recognition

Dimond (1958) examines the use of handwritten pattern recognition during the 1950s as a means of automating the verification of signatures by identifying forged handwriting, as well as recognising mathematical formulas, Roman, Chinese, Japanese and other handwritten scripts. The author further mentions techniques used in character recognition such as feature extraction, curve matching and time sequencing of zones.

Oliver Gordon Selfridge (1959) would publish a paper titled “Pandemonium: A Paradigm for Learning “, later to be used as a chapter in the book titled “Neurocomputing” (Anderson, Pellionisz and Rosenfeld, 1988). This book showcases the crossover and popularity between cognitive psychology and AI computing. Oliver would propose the first-ever computational model in pattern recognition known as the Pandemonium Architecture. The architecture would contain small components dubbed as “demons”, when put together, would allow machines to mimic the recognition of patterns by recording events and recognising patterns as they occur, triggering a subsequent event associated with that pattern.

Rosenblatt (1960) would publish a paper titled “Perceptron Simulation Experiments” introducing the brain model known as the Perceptron, the simplest version of this concept is represented by the diagram in Figure 1. It was aimed to mimic and model how the brain spontaneously recognises and classifies objects as well as process visual data. The experiment was conducted using photocells and random wiring, intended to recognise shapes such as letters. The author would conclude the simulation has a long way to go before it would become a feasible pattern-recognising device. Later that decade, Minsky and Papert (1969) would publish a book titled “Perceptrons” attacking Frank Rosenblatt’s work, suggesting it would be impossible for Perceptrons to learn an XOR function, the authors’ book discouraged research surrounding neural networks until the resurgence in the 1980s.

Figure 1: Rosenblatt’s Simple Perceptron Model (Rosenblatt, 1960)

2.2. CNN Architectures

Yann LeCun et al (1989) would publish a paper titled “Backpropagation Applied to Handwritten Zip Code Recognition” where the authors would propose Convolutional Neural Networks (CNN). With the use of back-propagation and minimal pre-processing the authors were able to create a fully automatic system to detect and classify 10 handwritten ZIP code numbers every second, out-performing manual designs. This paper would later be considered as the foundation for computer vision.

Typically, a CNN begins with feature extraction (performed by convolutional layers and subsampling) and ends with classification (performed by dense layers followed by a final Softmax layer). An example of a typical CNN design introduced by O'Shea, K. and Nash, R., (2015) uses two convolutional layers before each pooling layer, as illustrated in Figure 5. The increased number of convolutional layers resulted in more complex feature extraction.

Kaiming He et al (2016) published a paper titled “Deep Residual Learning for Image Recognition” introducing the Residual Neural Network (ResNet). The architecture would introduce the “identity shortcut connection” represented by the diagram in Figure 2. The identity shortcut connection allows the CNN to skip one or more layers depending on the identity of the input. The architecture would achieve an error rate of 3.57% outperforming humans on the dataset used (He et al. 2016).

Figure 2: Identity Shortcut Connection (He et al. 2016).

2.3. Conclusion

The literature suggests that the use of artificial intelligence reduces both resources and manpower, allowing for tasks such as identifying and sorting ZIP codes, forged signature recognition and processing of bank cheques to be fully automated. Despite Marvin Minsky’s (1969) scepticism on the limitations of AI, many examples of evolution and advancements

were given in this literature review, showcasing the potential of AI alongside the growth of computational power via Moore’s Law (Moore, 1965). Both reasons justify the need for techniques to be rigorously tested and rethought.

This raises several research questions concerning the use of artificial intelligence in handwritten digit recognition such as (RQ1) What is the relationship between filter sizes and classification accuracy? (RQ2) Which filter size should be recommended for the development of handwritten digit recognition using the MNIST dataset (LeCun and Cortes, 2010)? These questions will form the basis for the research to be carried out during this study.

Further research is needed to improve the robustness of digit recognition, examples such as using both limited and noisy data as training and test data are commonly overlooked and could pose a challenge to all types of variations in architecture. This research could lead to more sophisticated applications and use cases such as scanning and storing documents in real-time via webcam or assistive sentence recommendations using both pattern recognition and natural language processing (NLP) techniques.

3. Premise

Figure 3: Fishbone Diagram

Figure 3 illustrates the many variables that could affect the model’s classification accuracy. It would not be feasible to measure the influence of all the variables within the scope of this study. In this study, we’re isolating the filter size variable to explore its impact on classification accuracy.

To answer our research questions, this study uses the quantitative method of experimentation to explore two research objectives: (RO1) Evaluate the accuracy of convolutional filter sizes for the development of handwritten digits recognition (RO2) Recommend a filter size for handwritten digits recognition. To explore these objectives, a series of experiments will be conducted consisting of a standard CNN model, and the publicly available MNIST dataset (LeCun and Cortes, 2010). A minimum filter size of 1x1 will be applied and gradually increased up to 5x5, each time the filter is increased, the model's accuracy is recorded. The experiment is intended to explore whether increasing the filter size has any impact on the model's accuracy.

Accuracy is defined by the percentage of correctly predicted and classified digits in the testing data. If the testing data were to contain 100 digits from 0-9 and 90 out of 100 digits were correctly classified by the model, the model's accuracy would be 90%.

Using the research questions and our definition of accuracy, two hypotheses are created: H_0 : There is no significant difference between filter sizes and the model's $\bar{x}$ accuracy where $P > 0.05$. H_a : There is a significant difference between filter sizes and the model's $\bar{x}$ accuracy where $P \leq 0.05$. Statistically, we are trying to prove that increasing filter size has significance on the CNNs classification accuracy and reject the null hypothesis with a 95% level of confidence.

$$\bar{x} = \frac{\sum x}{n}$$

The $\bar{x}$ value refers to the mean classification accuracy where $\sum x$ = sum of all data values and n = number of items in the sample.

$$CV = \frac{\sigma}{\mu}$$

The CV value refers to the coefficient variation where σ = standard deviation and μ = mean accuracy. A $CV \geq 1$ indicates a relatively high variation whereas a $CV < 1$ indicates a low variation from the mean.

For details on legal and ethical considerations, please refer to section 8 and 9.

4. Methodology

4.1. Environment Setup

All experiments were conducted on a personal computer using the following:

Table 1: Software Versions Used

Version	Python Version	Compiler	cuDNN	CUDA
TensorFlow 2.4.0	3.7.0	MSVC 2019	8.0	11.0

4.2. MNIST Dataset

Figure 4: 28x28 Pixel Digit from MNIST

The MNIST dataset (LeCun and Cortes, 2010) is imported and accessed through the deep learning API Keras. The training set contains 60,000 28x28 pixel grayscale images of handwritten digits between 0 and 9, along with a test set containing 10,000 images. The digits have been size-normalized and centred in a fixed-size image; this is illustrated in Figure 4.

Table 2: Distribution of Digits in MNIST Dataset

Digit	0	1	2	3	4	5	6	7	8	9	Total
Train	5,923	6,742	5,958	6,131	5,842	5,421	5,918	6,265	5,851	5,949	60,000
Test	980	1,135	1,032	1,010	982	892	958	1,028	974	1,009	10,000

4.3. Network Architecture

Figure 5: CNN Architecture

Referring to the literature review, a common CNN design introduced by O'Shea, K. and Nash, R., (2015) uses two convolutional layers before each pooling layer, as illustrated in Figure 5. The authors strongly recommended the design due to the increased number of convolutional layers resulted in more complex and accurate feature extraction hence it will be used as a baseline for this study.

Feature Extraction:

The first two convolutional layers will consist of 32 layers. The second two convolutional layers will consist of 64. Initialises refer to how random weights are initially set. He uniform variance scaling initializer is applied to the convolutional layers along with the activation function ReLU. For every two convolutional layers, a max-pooling size of 2x2 will be applied as illustrated in Figure 5. $N \times N$ will be applied to all filter sizes where $N = 1$ and increments by 1 up to a filter size of 5x5.

Classification:

The fully connected layer contains a flatten layer, and two final dense layers. The first dense layer is sized at 100 using the He uniform variance scaling initializer and ReLU and the final dense layer is sized at 10 using Softmax.

4.4. Pre-Processing

Each image included in the dataset provided by MNIST (LeCun and Cortes, 2010) is categorised in 10 classes, between the digits 0 and 9. Machine learning algorithms only accept numerical inputs, hence the need to encode these categorical variables into numerical values

using encoding techniques (Potdar, S. and D., 2017). To solve this, one-hot encoding is used to transform the integers into a 10-element binary vector, 1 being the index class value and 0 for all other classes.

The pixel values for each image in the dataset contain an integer value between 0 and 255 (black and white). The pixel values are normalised by rescaling them between the ranges of 0 and 1, this is done by converting the pixel values from integers to floats and dividing them by the maximum value for a single colour channel which is 255.

4.5. Training and Testing Data

Training:

The CNN model has been trained for 10 epochs to ensure the experiments do not take too long as this individual study is an IS20 where limited time is applied. A standard batch size of 32 is also used in this experiment.

The training begins with the input layer holding the pixel value of the image (28x28), the input then goes through the first convolutional layer where a filter size of $N \times N$ is repeatedly applied to the input, overlapping its pixel values and extracting values onto a feature map. That feature map is then applied to the second convolutional layer which goes through a similar process. The feature maps are then applied to the 2x2 max-pooling layer, where the max pixel value from the filter is taken into the pool. The last two convolutional layers and max-pooling layer repeat the same process, extracting features from the previous layers. Once the features are extracted, the values then go through a fully-connected layer (FC) where the data is converted into a 1-dimensional array (flattened) and goes through two dense layers (neural networks) where the model will attempt to classify the image as a digit between 0-9.

Testing:

Using the train-test split method, the data is split, 83.4% ($N = 60,000$) of the data is used for training and 16.6% ($N = 10,000$) of the data is used for testing. This is used to avoid bias, ensuring the model is tested on unseen data. Cross-validation allows us to split our data into smaller equal-size groups. Using K-Fold cross-validation where $k = \text{number of folds}$, 5 folds are used, each fold containing a new shuffled sample for training and testing. The performance of the model can be calculated using the mean performance across all 5 k-folds (Problem: Cross-Validation – Machine Learning (Theory), 2021).

5. Findings

5.1. Experiment Results

Control Condition:

The control condition consists of the model described in section 4.3 with a filter size of 1x1. Unavoidable randomness such as weight initialisation was considered, hence why the experiment was repeated 5 times to minimise bias with further measures such as the train-test split method and k-fold cross-validation used to further minimise any anomalies found during training. 94.69% mean accuracy was derived from 5 experimental iterations, this will be used as a baseline for performance.

Experimental Condition:

The experimental condition was conducted using a three-stage process consisting of (1) increment the filter size $n \times n$ by one, (2) repeat the experiment 5 times and (3) calculate the $\bar{x}$ accuracy derived from the 5 repeated experiments. These 3 steps were repeated until a filter size of 5x5 was reached.

Table 3: Descriptive Statistics on Filter Sizes

Filter Size	Mean	Std. Deviation	CV	N
1x1	94.6994	.17588	0.00185	5
2x2	98.9166	.03221	0.00032	5
3x3	99.0174	.06024	0.00060	5
4x4	98.9232	.09085	0.00091	5
5x5	99.0034	.05685	0.00057	5
Total	98.1120	1.74423	0.01777	25

Table 3 illustrates the results recorded from the experiment (see Appendix F for raw data). The value N refers to the number of experiments conducted per filter size and ‘Mean’ refers to the mean accuracy derived from their 5 repeated experiments. The table also illustrates the standard deviation between each filter size. Using the CV formulae presented in section 3, the CV for filter sizes were calculated as shown in Table 3.

Table 4: Test of Between-Subject Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	72.828 ^a	4	18.207	1933.465	.000
Intercept	240649.114	1	240649.114	25555350.513	.000
Filter Size	72.828	4	18.207	1933.465	.000
Error	.188	20	.009		
Total	240722.130	25			
Corrected Total	73.016	24			

Table 4 illustrates the inferential ANVOA results showcasing the interaction between filter sizes and classification accuracy $F(1933.465) = 0.00$ where $P \leq 0.05$. These results lead us to accept the alternative hypothesis H_a .

5.2. Estimated Marginal Means

Table 5: Estimated Mean of Filter Sizes

Filter Size	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1x1	94.699	.043	94.609	94.790
2x2	98.917	.043	98.826	99.007
3x3	99.017	.043	98.927	99.108
4x4	98.923	.043	98.833	99.014
5x5	99.003	.043	98.913	99.094

Table 5 displays similar results to Table 3; however, Table 5 also illustrates the lower bound (lowest accuracy derived from 5 experiments) and upper bound (highest accuracy derived from 5 experiments) for each filter size.

Figure 6: Mean Accuracy by Filter Size Line Chart

Figure 6 visually illustrates the significant difference in classification accuracy between filter size 1x1 and 2x2 where 1x1 = 94.699% and 2x2 = 98.917%.

5.3. Post Hoc Tests

Table 6: Multiple Filter Size Comparison

Multiple Comparisons

Dependent Variable: Accuracy

Tukey HSD

(I) Filter Size	(J) Filter Size	Mean Difference		Sig.	95% Confidence Interval	
		(I-J)	Std. Error		Lower Bound	Upper Bound
1x1	2x2	-4.21720*	.06137	.000	-4.4009	-4.0335
	3x3	-4.31800*	.06137	.000	-4.5017	-4.1343
	4x4	-4.22380*	.06137	.000	-4.4075	-4.0401
	5x5	-4.30400*	.06137	.000	-4.4877	-4.1203
2x2	1x1	4.21720*	.06137	.000	4.0335	4.4009
	3x3	-.10080	.06137	.489	-.2845	.0829
	4x4	-.00660	.06137	1.000	-.1903	.1771
	5x5	-.08680	.06137	.626	-.2705	.0969
3x3	1x1	4.31800*	.06137	.000	4.1343	4.5017
	2x2	.10080	.06137	.489	-.0829	.2845
	4x4	.09420	.06137	.553	-.0895	.2779
	5x5	.01400	.06137	.999	-.1697	.1977
4x4	1x1	4.22380*	.06137	.000	4.0401	4.4075
	2x2	.00660	.06137	1.000	-.1771	.1903
	3x3	-.09420	.06137	.553	-.2779	.0895
	5x5	-.08020	.06137	.690	-.2639	.1035
5x5	1x1	4.30400*	.06137	.000	4.1203	4.4877
	2x2	.08680	.06137	.626	-.0969	.2705
	3x3	-.01400	.06137	.999	-.1977	.1697
	4x4	.08020	.06137	.690	-.1035	.2639

*. The mean difference is significant at the 0.05 level.

Table 6 illustrates all the possible pairwise comparisons for the five groups and the associated p values. Table 7 suggests that there's only significance when comparing filter size 1x1 to larger filter sizes.

Table 7: Homogeneous Subsets

Tukey HSD^{a,b}

Filter Size	N	Subset	
		1	2
1x1	5	94.6994	
2x2	5		98.9166
4x4	5		98.9232
5x5	5		99.0034
3x3	5		99.0174
Sig.		1.000	.489

Table 7 illustrates two subsets, grouped by similar means. Subset 2 has a P-value of .489 where $P > 0.05$ suggesting there's no association between subset 2 and the model's classification accuracy.

5.4. Tests for Normality

A normality test is a statistical test to determine whether a dataset deviates from the expectations of a normal distribution. Two possible hypotheses are created during this test; H_0 : The values are sampled from a population that follows a normal distribution where $P > 0.05$. H_a : The values are not sampled from a population that follows a normal distribution where $P \leq 0.05$. To test these hypotheses, the Kolmogorov-Smirnov and Shapiro-Wilk test will be conducted on the data collected.

Table 8: Normality Test

	Filter Size	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Accuracy	1x1	.197	5	.200*	.951	5	.743
	2x2	.219	5	.200*	.957	5	.786
	3x3	.239	5	.200*	.905	5	.437
	4x4	.234	5	.200*	.893	5	.375
	5x5	.235	5	.200*	.848	5	.190

Table 8 displays the P-Value derived from the Kolmogorov-Smirnov and Shapiro-Wilk test. The results from both tests show that $P > 0.05$ allowing us to reject H_a : The values are not sampled from a population that follows a normal distribution based on a 95% level of confidence and accept H_0 : The values are sampled from a population that follows a normal distribution with 95% level of confidence.

6. Evaluation

Using the statistical analysis conducted in section 5, there appears to be some association between the filter sizes and the model's classification accuracy. When increasing the filter size from 1x1 to 2x2, significance was shown where $P = .000$.

However, Table 7 shows us that filter sizes between the ranges of 2x2 to 5x5 had a P-Value of .489 where $P > 0.05$. This shows us that although there is significance when a filter size of 1x1 is increased to 2x2, there isn't enough significance to recommend increasing a filter size larger than 2x2 for the development of handwritten digit recognition using MNIST (LeCun and Cortes, 2010).

The results in Table 3 indicate the CV for all filter sizes were below 1 suggesting little variation from the mean classification accuracy. If their results were above 1, it would have indicated more data needed.

We can reject H_0 : There is no significant difference between filter sizes and the model's $\bar{x}$ accuracy based on a 95% level of confidence and accept H_a : There is a significant difference between filter sizes and the model's $\bar{x}$ accuracy with a 95% level of confidence.

Due to the experiment only being conducted on one dataset, more data is needed to evaluate the association between filter sizes and classification accuracy by conducting a similar experiment using varied datasets.

7. Discussion

This study was set out to provide a chronological overview of pattern recognition techniques using handwriting recognition as well as an in-depth analysis of CNN's. While conducting the literature review, the number of papers on the association between filter sizes and classification were limited and none had explored the association using handwritten digit datasets. This knowledge gap drove this study to test the association between filter sizes and classification accuracy using a handwritten digit dataset such as MNIST (LeCun and Cortes, 2010).

Initially, I had intended to conduct a comparative analysis on various CNN architectures but after further research on the number of variables that could potentially be affecting the model's classification accuracy, I had decided it would be unfeasible and out of scope to test the many variables illustrated in Figure 3. Filter sizes were isolated for this study due to the

limited research on its association with classification accuracy using handwriting datasets. As a result, the literature review conducted in this study may have become less relevant to the methodology and findings presented. More research needs to be conducted on the existing literature on the association between filter sizes and classification accuracy but due to the limited time given for an IS20, it would have been unfeasible to go back and re-conduct the literature review.

The CNN model used for testing was a common CNN design introduced by O'Shea, K. and Nash, R., (2015). It was recommended due to its increased number of convolutional layers resulting in more complex and accurate feature extraction. The statistical method ANOVA was used in this study to statistically determine significance between the means of 5 groups (filter sizes), ranging from 1x1 to 5x5. This method of analysis was chosen over methods such as T-tests due to the results containing more than two groups.

Our findings indicate that increasing the filter size from 1x1 to 2x2 can impact the model's classification accuracy on the MNIST dataset. The results suggest no significance when increasing the filter size above 2x2. As a result, a filter size above 2x2 could not be recommended for the development of handwritten digit recognition. These findings suggest further exploration is needed on various filter sizes, CNN models and datasets to test whether the results found in this study are robust.

This research makes two contributions. First, it extends the literature on the association between filter sizes and classification accuracy. Second, a recommendation of a filter size 2x2 is given for the development of handwritten digit recognition using MNIST.

8. Legal Considerations

8.1. Health and Safety at Work Act 1974

The Health and Safety at Work Act 1974 requires staff to be trained to ensure health and safety procedures are understood, as well as a properly maintained and safe environment in the workplace (Health and Safety at Work etc. Act 1974, 2021). No activities conducted in this study require health and safety training nor does it require supervision.

8.2. Computer Misuse Act 1990

The Computer Misuse Act 1990 refers to the unauthorised access to computer materials, unauthorised access with intent of committing or aiding further offences and unauthorised modification of computer material (Computer Misuse Act 1990, 2021). No activities of this nature will be conducted during this study.

8.3. Data Protection Act 1998/2018

This study does not involve the collection of any data from human participants, any data being used, is from the publicly available MNIST dataset (LeCun and Cortes, 2010), therefore there are no legal considerations needed to be made regarding the Data Protection Act 1998/2018 or General Data Protection Regulation (GDPR) (Data Protection Act 2018, 2021).

9. Ethical Considerations

9.1. Ethics Review Checklist

An “**Ethics Review Checklist for Undergraduates**” document has been filled out and submitted to the supervisor of this study, Dr Richard Henson (see Appendix G). The ethical issues listed and proposed on the document had no relation to this study and the document was approved shortly after by the supervisor during this study.

9.2. Unethical Future Research

This study is intended to conduct a correlational analysis on the association between filter sizes and classification accuracy. This could potentially lead to unethical research such as creating an algorithm so sophisticated and accurate resulting in forged handwriting undistinguishable by the human eye. However, this paper heavily discourages any misconduct and research of the nature, due to the unethical and legal implications surrounding identity theft. This paper is solely intended to explore the correlation between filter sizes and classification accuracy.

10. Conclusion

The original intention for this dissertation was to conduct a comparative analysis on various CNN architectures on handwritten digit recognition however when exploring the many variables that could affect a model's classification accuracy, 14 variables were found, refer to Figure 3. Exploring 14 variables in an IS20 would have been unfeasible and out of scope, hence why the focus of this study was concentrated on one variable.

Two research questions were set out to be answered during this study; (RQ1) What is the relationship between filter sizes and classification accuracy? RQ1 Answer: The findings suggest there is an association between filter sizes and the model's classification accuracy but above a size of 2x2, filter sizes become less significant. (RQ2) Which filter size should be recommended for the development of handwritten digit recognition using the MNIST dataset? RQ2 Answer: A filter size of 2x2 is recommended for the development of handwritten digit recognition using MNIST due to its significantly increased accuracy of 98.91% as opposed to 94.69% from a filter size of 1x1. A filter size above 2x2 cannot be confidentially recommended due to its insignificance.

Regarding limitations, as previously mentioned, my original intention was to study variations of CNN architectures but due to the scope of the study, there was not enough time given to explore more than 1 CNN variable. Had the deadline been increased, more variables would have been explored, allowing us to conduct a multi-variable analysis. It should also be mentioned, a lot of time would have been saved if the topic of my research proposal hadn't needed to change. Regarding our findings, there is not enough data to conclude whether the correlation between filter sizes and classification accuracy implies any causation, more research needs to be done to explore the causation on CNN's classification accuracy as well as its space complexity.

Our results were solely measured using accuracy, this proffers an opportunity to explore the association between filter sizes and space complexity. The results would allow scholars to confidently recommend a filter size for the development of handwritten digit recognition. This could be done using varied sizes of training datasets and model depth.

References

- Ahmed, W. and Karim, A., 2020. The Impact of Filter Size and Number of Filters on Classification Accuracy in CNN. 2020 International Conference on Computer Science and Software Engineering (CSASE),.
- Brownlee, J., 2021. How to Develop a CNN for MNIST Handwritten Digit Classification. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/how-to-develop-a-convolutional-neural-network-from-scratch-for-mnist-handwritten-digit-classification/> [Accessed 22 April 2021].
- Brownlee, J., 2021. How To Improve Deep Learning Performance. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/improve-deep-learning-performance/> [Accessed 5 April 2021].
- Dimond, T., 1958. Devices for reading handwritten characters. Papers and discussions presented at the December 9-13, 1957, eastern joint computer conference: Computers with deadlines to meet on XX - IRE-ACM-AIEE '57 (Eastern).
- GitHub. 2021. sambit9238/Deep-Learning. [online] Available at: https://github.com/sambit9238/Deep-Learning/blob/master/cnn_mnist.ipynb [Accessed 22 April 2021].
- Hammond, K., 2015. Practical Artificial Intelligence for Dummies. John Wiley & Sons, Inc.
- He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
- Hunch.net. 2021. Problem: Cross Validation – Machine Learning (Theory). [online] Available at: <https://hunch.net/?p=29> [Accessed 22 April 2021].
- Krizhevsky, A., Sutskever, I. and Hinton, G., 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), pp.84-90.
- LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W. and Jackel, L., 1989. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation, 1(4), pp.541-551.

- Leedham, C.G., 1994, March. Historical perspectives of handwriting recognition systems. In IEE Colloquium on Handwriting and Pen-based input (pp. 1-1). IET.
- Legislation.gov.uk. 2021. Computer Misuse Act 1990. [online] Available at: <https://www.legislation.gov.uk/ukpga/1990/18/contents> [Accessed 22 April 2021].
- Legislation.gov.uk. 2021. Data Protection Act 2018. [online] Available at: <https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted/data.htm> [Accessed 22 April 2021].
- Legislation.gov.uk. 2021. Health and Safety at Work etc. Act 1974. [online] Available at: <https://www.legislation.gov.uk/ukpga/1974/37/contents> [Accessed 22 April 2021].
- MATLAB, 2020. version 9.9.0.1570001 (R2020b) Update 4, Natick, Massachusetts: The MathWorks Inc.
- Michal, Kozielski, Doetsch, P. and Ney, H., 2013. Improvements in RWTH's System for Off-Line Handwriting Recognition. 2013 12th International Conference on Document Analysis and Recognition,.
- Mudialba, P., 2016. The Impact of Cloud Technology on the Automation of Businesses. 2016 International Conference on Platform Technology and Service (PlatCon).
- O'Shea, K. and Nash, R., 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.
- Pham, H., Setlur, A., Dingliwal, S., Lin, T., Poczos, B., Huang, K., Li, Z., Lim, J., McCormack, C. and Vu, T., 2020. Robust Handwriting Recognition with Limited and Noisy Data. 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR),.
- Potdar, K., S., T. and D., C., 2017. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. International Journal of Computer Applications, 175(4), pp.7-9.
- Rosenblatt, F., 1960. Perceptron Simulation Experiments. Proceedings of the IRE, 48(3), pp.301-309.
- Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

UK broadband usage more than doubled in 2020. [online] Available at:

<https://www.openreach.com/news/uk-broadband-usage-more-than-doubled-in-2020---driven-by-live-sport-online-gaming-and-home-working/> [Accessed 10 February 2021].

Yann.lecun.com. 2021. MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges. [online] Available at: <http://yann.lecun.com/exdb/mnist/> [Accessed 31 January 2021].

Bibliography

Anderson, J., Pellionisz, A. and Rosenfeld, E., 1988. Neurocomputing. Cambridge, Mass.: MIT Press, pp.115–122.

Minsky, M. and Papert, S., 1969. Perceptrons. Cambridge, Mass: Massachusetts Institute of Technology.

Appendix A: Glossary

AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
HWR	Handwriting Recognition
SNARC	Stochastic Neural Analog Reinforcement Calculator
MNIST	Modified National Institute of Standards and Technology
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
ResNet	Residual Neural Network
Epoch	Refers to one cycle through the full training dataset
FC	Fully Connected Layer
One-Hot Encoding	Integer encoded variable is removed and a new binary variable is added for each unique integer value