National College of Ireland

Improving Image Classification Performance Through Advanced Deep Ensemble Techniques

MSc Research Project
Data Analytics

Pratik Anil Dhaygude
Student ID: x22108670

School of Computing
National College of Ireland

Supervisor: Barry Haycock

National College of Ireland
MSc Project Submission Sheet
School of Computing

Student Name:	Pratik Anil Dhaygude
Student ID:	x22108670
Programme:	MSc Data Analytics **Year:** 2024
Module:	MSc Research Project
Supervisor:	Barry Haycock
Submission Due Date:	12th August 2024
Project Title:	Improving Image Classification Performance Through Advanced Deep Ensemble Techniques
Word Count:	…11088…… **Page Count** …31……..

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: *Pratik dhaygude*

Date: 12/08/2024

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):	

Improving Image Classification Performance Through Advanced Deep Ensemble Techniques

Pratik Anil Dhaygude
x22108670

Abstract

Improvement of image classification forms the core of the study with refined deep ensemble strategies. Chapter 1 introduces by defining the problem, aim and objectives are defined as applied in enhancing the image classification accuracy, robustness, and explainability with the utilization of ensemble approaches.
Chapter two provides the literature review of image classification and deep ensemble, stacking, bagging, boosting and hybrid methods.
Methodology is described in chapter three with an experimental design selected to facilitate the achievement of the research objectives. It outlines ways of generating and deploying the ensemble models as well as data pre-processing, model designs and training algorithms
Chapter four provides an overview of the process of applying and assessing the outcomes of employing different machine learning techniques for image classification using the "Flower" dataset. During EDA, class distributions and characteristics of images are found and the dataset is made more ready by performing data augmentation. Implementation of the model comes into CNN and some of the latest architecture such as ResNet 50, VGG 16, and Inception V3. Ensemble techniques used were stacking, Bagging, and Boosting. Other evaluation measures such as training and validation loss, accuracy, a confusion matrix, and classification report are given and the accuracy level varies from 68% to 94%. The presented results allow identifying the success rates of various models and techniques in image classification.
Chapter 5 compares the performance of different machine learning algorithms when applied to the ''Flower'' dataset. CNN had the highest accuracy of 68% but the model seemed to overfit, and therefore regularization is needed. For feature extraction, the pre-trained models including ResNet50, VGG16, and InceptionV3 expressed good results. Thus, Stacking, Bagging, and Boosting were used and comparing the results we can state that Bagging has higher precision and F1-score. the percentage of classification accuracy remained rather moderate.
The last chapter demonstrated that ensemble models have helped enhance learners' image classification accuracy. These objectives were met by creating and fine-tuning these techniques to show that they can indeed improve the model's stability and performance. It has been reported that future advancements should also use more significant datasets, develop efficient computational strategies for effective execution of the models and enhance the readability of the same for better referencing or application.

Chapter 1: Introduction

1.1 Background and overview

Image classification is one of the basic types of tasks in computer vision and deals with the correct classification of images into certain classes. Still overfitting problems, local optima problems and the scarcity of labelled data are existing problems in images which cannot be overlooked despite the incredible improvement in the field of deep learning. Goal of this research is to improve the image classification effectiveness by using deeper ensembling techniques such as stacking, bagging and boosting involve the making of several models and bringing them together to increase the reliability and precision of images. It is in this context that the study seeks to propose and put into practice the mentioned strategies with a view of solving some inherent challenges in deep learning training and in particular increasing efficiency as best as possible.

Suggested techniques can be compared quantitatively to traditional approaches via standard image datasets accuracy, precision, recall, F1-score, and computational complexity can be used for comparison. This research outcomes can help enhance the existing computer-aided diagnosis systems especially in medical image analysis by enhancing the reliability and interpretability of classification models (Abimannan *et al.* 2023). This research endeavours to contribute to the advancement of image classification technology by providing novel ensemble methods and valuable insights into their performance and applicability across diverse datasets.

1.2 Research aim and objectives

Aim

Aim of this research is to enhance the performance of image classification by utilising advanced deep learning ensemble techniques.

Objective

- To develop and implement ensemble strategies such as stacking, bagging, boosting, and hybrid methods
- To investigate different optimisation algorithms and techniques for efficient training and deployment of ensemble models
- To evaluate the performance of the proposed ensemble techniques using standard image classification datasets
- To compare the performance of the ensemble methods using performance matrices such as accuracy, precision, recall, F1-score, and computational efficiency

1.3 Research question

1. Which ensemble configurations work best to maximize the accuracy of image categorization across a variety of datasets?
2. Which deep ensemble technique produces the greatest gain in robustness for image classification problems?
3. Which techniques for incorporating deep ensemble approaches into image classification models effectively reduce overfitting?
4. How does the interpretability of picture categorisation findings change when ensemble learning is included?

1.4 Research rationale

Following research can address the fluctuating and leading issues in image classification such as over-fitting, local optima and the lack of tagged images. It is also expected that applying several sophisticated deep ensemble methods such as stacking, bagging, and boosting helps to improve the model's robustness, accuracy, and readability can be achieved. These techniques are used to inherent training issues and enhance model performance as well as the classified system's

robustness (Silva *et al.* 2020). Proper classification methods are considered for an image which is the primary field of study and the objectives of the research are to contribute to advancements in image visualisation systems and offer improved image processing technologies incorporating enhanced interpretability that in turn can result in the optimization of image classification.

1.5 Problem statement

There are some significant open problems in image classification such as overfitting, local optima, and restricted labelled data where deep learning has been applied to reduce these problems. These issues significantly reduce the efficiency and reliability of the developed classification models, exceptionally while operating in areas such as medical imaging. These are called overfitting and they lead to a bad outcome with the data not used for creating models while local optima make the models remain in low-performing states. This is made worse by the availability of few labelled datasets mostly due to the inability of the model to learn good and separable features.

This research aims to reduce several challenges by using deeper ensembling techniques including stacking, bagging, and boosting. This deep learning is a compilation of several classification models so the advantages of the several performances are taken to improve the quality, reliability, and clarity of images (Younas *et al.* 2023). The integration of the above-mentioned techniques into the deep learning frameworks proved to be more challenging because of the nesting of staking operation which is an intrinsic property in deep neural networks and due to computational overhead in the sense that it requires training of the number of models simultaneously.

1.6 Significance of the research

This research is important as it allows to improve performance for recurrent difficulties in image classification which forms the basis of numerous applications in the computer vision domain. Deep learning ensemble methods have been hugely prominent problems from the images such as overfitting, and local optima, and again the limited availability of labelled data poses a significant threat to precisely accuracy-oriented classification models. These problems are more disastrous in delicate fields such as medical imaging where the accuracy of diagnosis is vital. Deep learning ensemble methods such as stacking, bagging, and boosting are applied in this work which improve the stability, accuracy, especially, interpretability of those classification models (Mienye *et al.* 2022). Deep learning challenges such as over-fitting, local optima and the lack of tagged images can be solved by ensemble methods, which make use of the features of a set of models that are integrated into a single model while limiting the drawbacks of individual models.

This study is aimed at improving the feasibility of the training and deployment of these complex methods to serve real-world solutions. The findings of this study apply to image classification systems in image classification methodology. There is a possibility of extending the findings of this study to contribute to the enhancement of the overall area of image classification while offering guidelines regarding the utilization of ensemble methods in different datasets and contexts (Ahmed *et al.* 2023). This study is useful for improving current and future research on image classification technologies that are reliable, precise, and easy to explain.

1.7 Research hypothesis

Null hypothesis (H0):

Application of advanced deep learning ensemble techniques such as stacking, bagging, and boosting does not significantly improve the performance (in terms of recall, F1-score, accuracy, precision, and computational efficiency) of image classification models compared to traditional approaches.

Alternative hypothesis (H1):

Application of advanced deep learning ensemble techniques such as bagging, stacking, and boosting significantly improves the performance (in terms of accuracy, precision, recall, F1-score, and computational efficiency) of image classification models compared to traditional approaches.

1.8 Research structure

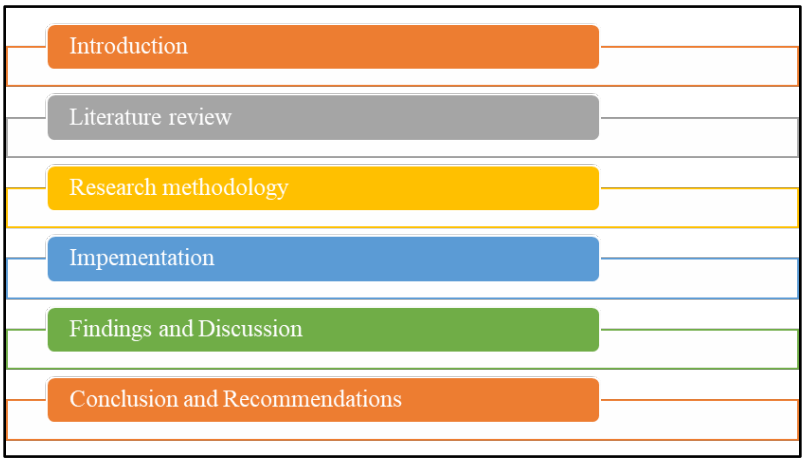

Figure 1.8.1: Research structure

1.9 Summary

This study discusses the various deep-learning ensemble methods that are used for classifying images by enhancing their performance such as precision, accuracy, F1-score, computational efficiency, and F1-score by using classified models. The introduction identifies the study on improving image classification through the application of deep learning ensemble methods such as stacking, bagging and boosting which solves issues such as overfitting, local optimum and the restriction of available labelled data. Potential applications of the convolutional neural networks or CNN are to increase the model's trustworthiness, precision, and readability in medical imaging. These include creating and deploying these ensemble techniques, fine-tuning the training, and deployment processes, and assessing the efficiency of the strategies based on customary measures. The value is based on enhancing the methods underlying image classification, providing some ideas regarding the usage of these approaches, and possibly contributing to the development of CAD systems.

Chapter 2: Literature review

2.1 Introduction

Image classification has been seen as a core concept in computer vision due to the development of the deep learning approach. The basic purpose here is to have the capability to classify images into a pre-specified set of classes, which has been explained to witness immense improvement with the help of convolutional neural networks (CNNs). The most common techniques by which the deficiencies of decision trees can be handled include ensemble methods that include stacking, bagging and boosting, among others. These methods use several models to get better results. These models have limitations when used individually, which is why this approach offers a better solution to the problem. This chapter compares contemporary works on image classification and ensemble strategies regarding their progress, advantages, and disadvantages. This also explores the theoretical framework and implementation history of these techniques' writings, including achievements and future directions. Thus, the section has positioned itself to offer a brief review of the state of the image classification before using ensemble techniques and possibilities of enlarging the model's accuracy, robustness and interpretability.

2.2 Themes

2.2.1 Ensemble method strategies in Image classification

Image classification has encouraged the use of ensemble methods because they improve the model's accuracy, reliability, and malleability. These methods use several classifiers and attempt to use each of their strong points while simultaneously avoiding the pitfalls that come with them. The studies of ensemble techniques in image classification identify some basic methods such as bagging, boosting, stacking, and others that are a combination of the basic ones that help to increase the performance in various tasks. As per the view of Szmurło, and Osowski, 2022, previous techniques used to classify images included handcrafted features and fewer classifiers. HOG, SIFT, and PCA were some of the common methods used for feature extraction. These features were then supplied to classifiers such as the Support Vector Machines (SVMs) or the decision trees. Even though the listed methods provide moderate results, the quality of the handcrafted features and the complexity of the image data negatively affect the result.

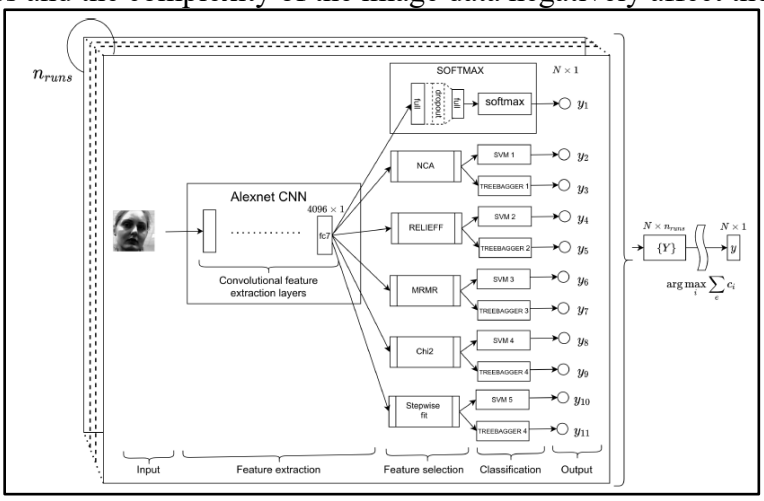

Figure 2.2.1.1: CNN Layer architecture

(Source: Szmurło, and Osowski, 2022)

New approaches, such as deep learning, such as CNN, have changed the approach to image classification. CNNs do not require manual feature extraction since these are endowed with the characteristic of learning from raw image data and hierarchical features. The progress on many image classification tasks has shown that architectures such as AlexNet, VGG, ResNet, and Inception are highly efficient (Li et al. 2023). AlexNet recorded an impressive performance in the ImageNet challenge with the help of deep structures and big data sets. Data augmentation strategies have been used significantly. Operations such as flips, rotations, translations, and adding noise assist in constructing models to other data, making the resultant models capable of generalising from the original training data set to other datasets. Additional ways of Regularisation, such as Dropout, weight decay and Normalisation using batches, also help improve generalisation by reducing overfitting.

Ensemble methods combine multiple models to enhance the classification results and the algorithms' stability. Originally, bagging, boosting, and other machine learning techniques like stacking have generated great diverse and independent classifiers, increasing accuracy. In the same context, improved performance has been noted when different classifiers such as the SVM, the decision trees, and soft-max classifiers have been used in parallel through majority votes or weighted averages. Transfer learning applies a concept trained on one large dataset and another on another specific data set (Zheng et al. 2023). This approach comes in handy when the target dataset

is small, meaning that the model can rely on the features learned by the first model. Features learned on ImageNet can be transferred to the facial recognition task, which is likely to improve the performance. Face recognition is an example of an image classification application with a vast amount of research work done to enhance the system's performance. Recent methods are mostly based on deep learning approaches. By leveraging data augmentation and regularisation and using CNN-based architectures with ensemble methods, facial recognition tasks attain significant performance.

2.2.2 Deep learning Ensembling method techniques

The classification of brain tumours based on MRI images is one of the most important subfields in medical image analysis because of the urgent need for better and faster diagnosis. Current literature reveals interesting progress and issues in this area of research. Research done in earlier years mainly focused on applying classical methods belonging to the field of machine learning for the classification of brain tumours. SVM, k-NN, and RF were applied in categorising tumours by using features that were obtained from the MR images. This research united wavelet-based comparison of GLCM characteristics and OFPA opposition for exclusion. As per the view of Kang et al. 2021, the study achieved 92% accuracy as it portrayed some of the added features those traditional methods possessed but at the same time failed to address fully, such as non-linear and multi-dataset classes.

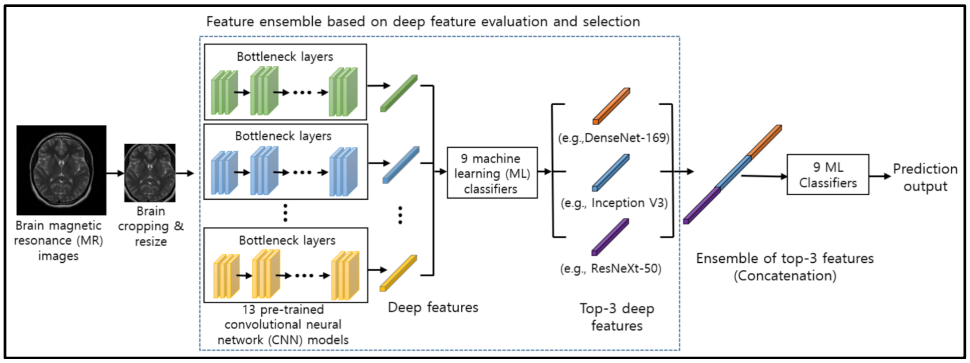

Figure 2.2.2.1: Feature ensemble based on deep feature evaluation and selection

(Source: Kang *et al.* 2021)

The evolution to deep learning opens a new advancement in the classification of brain tumours. CNN can learn features at multiple levels directly from the image and enhance the classification. This led various research to emphasise that the pre-trained CNN such as ResNet, DenseNet and Inception are well suited to extract the desired feature from the MRI images of the brain. The results reveal that image processing, along with the probabilistic neural network (PNN), is used for tumour detection with moderate accuracy and less computational time (Afify *et al.* 2020). This goes a long way into illustrating how deep learning is eloquent in feature extraction and classification. Although deep learning models exist and are used in many applications, deep learning models encounter difficulties. A final disadvantage of deep learning in medical imaging is that there are often limited large annotated datasets.

Most of the studies are concerned with normal and abnormal classification, with little mention of multi-classification, which is very essential when differentiating between different types of tumours. This study aims at investigating the usefulness of Transfer Learning (TL) and Ensemble Learning (EL) methodologies in classification of images and put forward the Ensembled Transfer Learning (ETL) framework (Rai and Pahuja, 2024). This limitation is resolved by the recent development including the current study where an ensemble system of features from two different pre-trained CNN models is suggested. This method to improve the peaks and building blocks' top

features must integrate various machine learning classifiers to boost the classification ratio's precision and strength. Incorporating transfer learning has also been important as well. Fine-tuning of deep models pre-trained on other large databases such as ImageNet greatly cuts down the model training time and the amount of data required for the classification of brain MRIs. This approach has been supported by different studies stating higher classification success rates with the help of small datasets of MRI.

2.2.3 Transfer and Ensemble learning techniques for Cervical image classification

Cervical cancer is among the common cancer-related diseases in the developing world, early detection of cervical cancer plays a significant role in the management of the disease. Histopathological image analysis is still the reference for diagnosis but the manual analysis by pathologists is subjective, time taking and varies in accuracy. As per the view of Xue e*t al.* 2020, the research focuses on the classification of cervical histopathological images into differentiation stages using a new ETL framework with the view of improving the results of diagnostics using modern and efficient machine learning algorithms more accurately.

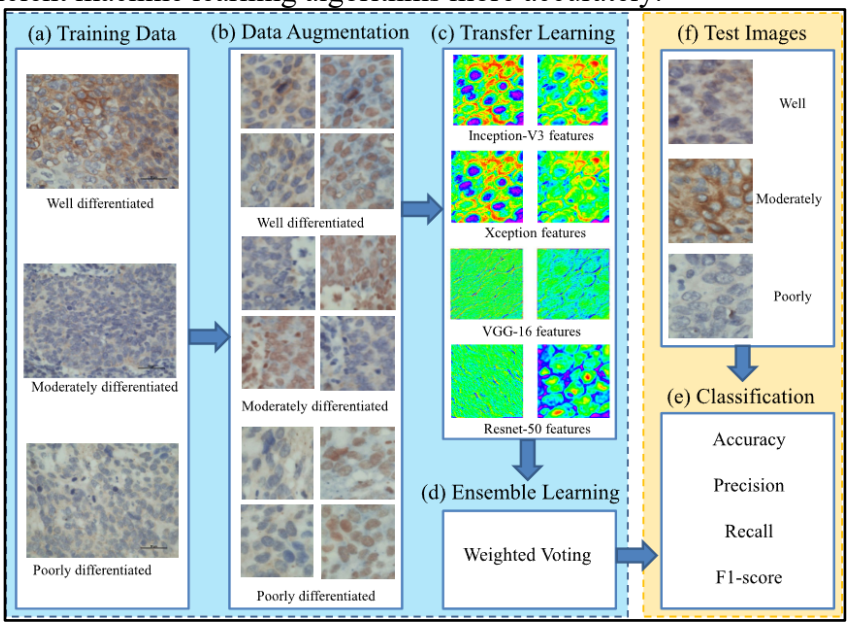

Figure 2.2.3.1: ETL framework

(Source: Xue e*t al.* 2020)

TL structures built in this research use Inception-V3, Xception, VGG-16, and Resnet-50 networks and present a weighted voting EL strategy that boosts classification. When assessing the outcomes on the image database consisting of 307 immunohistochemical stained images through AQP, HIF, and VEGF methods, the framework was found to have high accuracy of on AQP staining images. The immunohistochemical staining results show that AQP is expressed in the renal cortex and outer medulla, but not in the inner medulla or papilla (El-Hady *et al.* 2023). 61% where the images were poorly differentiated on VEGF staining. Other experiments performed on the Herlev dataset with the aim of distinguishing between benign and malignant cells concluded with an accuracy of 98.37%.

TL employs transfer learning of deep learning models such as Inception-V3, Xception, VGG16, and Resnet-50 which already have been trained on large databases including ImageNet. These models are employed for classifying cervical histopathological images where knowledge from generalised image recognition and analysis tasks are transferred to medical image analysis (Idlahcen *et al.* 2020). Other parameters are fine-tuned to enhance the accuracy for the given data;

the learning rate is lowered based on the training iterations. In order to enhance the distribution classification procedure, EL applies a weighted voting procedure which unites four different TL networks that allow the program to integrate numerous TL models. Every single network gives a set of category scores, and after that, the classification is done based on the sum of these scores, where the weights belonging to each network are different. For weighting, accuracy, recall, precision, and F1-score were used, out of which recall gave better results. The material used for the analysis consists of 307 cervical histopathological images that were stained by the AQP, HIF, and VEGF staining techniques. Images are grossly categorised as well, moderately, or poorly differentiated, and the images are boosted to improve the training efficiency.

	AQP	HIF	VEGF
Inception-V3	73.81	75.77	73.97
Xception	79.59	78.77	81.31
VGG-16	88.28	90.17	91.38
Resnet-50	90.88	91.86	90.69
Inception-V3 TL	77.50	79.27	77.62
Xception TL	84.08	82.78	85.62
VGG-16 TL	94.28	94.24	94.86
Resnet-50 TL	94.93	95.68	94.79

Figure 2.2.3.2: Accuracy of De-novo trained CNN and TL methods
(Source: Xue e*t al.* 2020)

Data table presented in figure 2.2.3.2 is done based on the accuracy, precision, recall, and F1-score, while 9-fold cross-validation helps in more realistic analysis. From the result analysis, it is clear that differentiation attains the highest accuracy, precision, and F1-score higher than all TL methods and the ETL framework, excluding recall, which recorded the corresponding best of moderate differentiation in all TL methods except for VGG-16. Application of VEGF staining always presents the highest classification prognosis and exactness in most of the TL methods and also in ETL framework. Comparing AQP staining with Inception-V3, Xception, and the ETL method, the highest value of recall is achieved. At the same time, for the model of HIF, the highest recall is obtained for Resnet-50. This study shows that the ETL method yielded a higher accuracy in the classification for all the differentiation stages and staining methods when compared to individual TL methods. The best overall accuracy is identified for poorly differentiated VEGF staining images with the value of 98.61%. Use of TL and EL techniques and the enhancement of the adopted ETL framework provide a better way to classify cervical histopathological images to the various differentiation stages and staining methods.

2.2.4 Performance of the ensemble learning methods in image classification

Ensemble methods or ensemble learning is the process by which from machine learning, higher and better performance and accuracy can be obtained by using combination techniques of different classification models. It combines the results of various models and improves the final output. Ensemble learning improves feature transformation, multi-layer features and deeper features are being extracted, uses different structure of models against this technique. Deep learning models are used to get whole performance matrices like precision, f1, recall and accuracy. Ensemble methods are used to fulfil the need of deep learning techniques faster and in an effective manner. Models such as CNN, RNN or another method help to compare the performance of the models where it is used. A reliable system can be developed using deep-ensembled method (Loddo *et al.* 2022). In CNN, feature extractions are dependent on convolution layers.

Other than that, the pooling layers are used for deducting the computational complexity. In this scenario, deep learning techniques like ensemble learning conduct various primary learners by using fusion techniques to upgrade whole performance. Structure of ensemble learning is more user friendly and it's also easy to understand. Main attraction of ensemble method is that it can use multiple CNN model by combine them. After using those, ensemble model can solve the error of one single model by using others. By doing so, overall performance is improved and provides better output that any single model (Ahad *et al.* 2023). Similar network structure with diverse model checkpoint can be defined with a specific training procedure. Also, identical network arrangement multiple training procedures can be used for the whole training process.

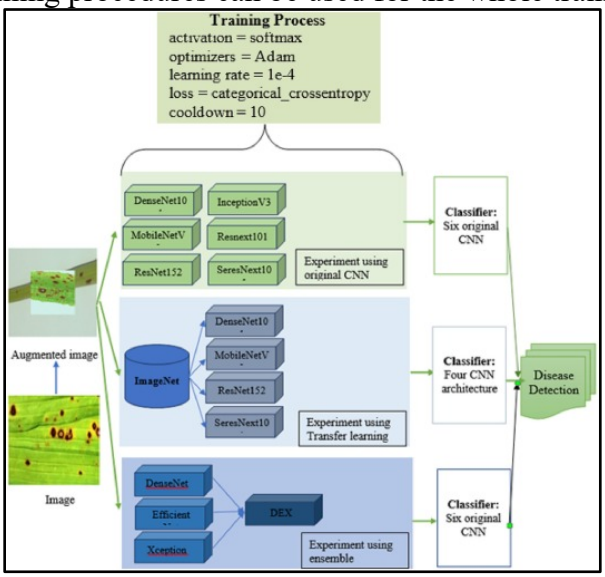

Figure 2.2.4.1: Training process of ensemble methods

(Source: Ahad *et al.* 2023)

Impact on ensemble methods makes the performance more accurate, performance of ensemble models is encouraged and linked with a test dataset. Then performance of models is evaluated by using testing procedures. Using multiple axes along with optimal window for the ensemble method produce better performance than any process using similar axis in ensemble methods. In this case, AUC and AIC curves are improved by using the ensemble methods of different axes. In this scenario, a decision can be taken that a better and effective performance is expected by using different axes and window settings (Nobashi *et al.* 2020).

Ensemble learning approach known as bagging involves constructing classifiers from different subsamples of the original dataset. But it is disadvantageous in the aspect that the procedure of classification is based on the random selection of training objects. Enhanced Bagging (eBagging) which is another modification in the bagging process and is developed to solve this problem by incorporating error-based boot-strapping in the process of preparing the training set (Tüysüzoğlu *et al.* 2020). After this the prediction and performance matrices are evaluated. This shows that when it comes to the classification of data points as well as controlling the training error, eBagging outperforms the rest of the techniques.

2.3 Conceptual framework

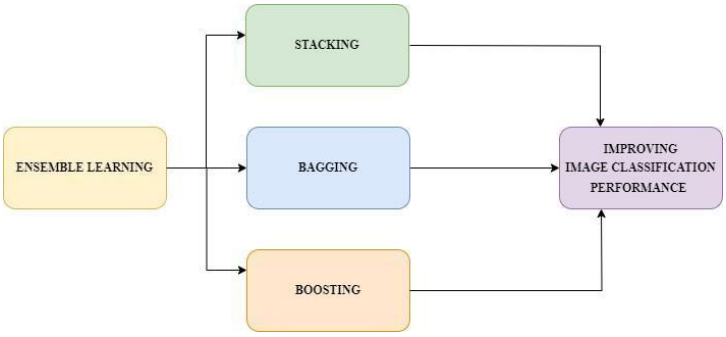

Figure 2.3.1: Conceptual framework

Here, Ensemble learning methods inclining Stacking, Bagging and Boosting is the independent variable. On the other hand, improving image classification performance is the dependent variable. Meaning performance improvement of image classification relies on the successful integration of ensemble learning methods.

2.4 Literature gap

A number of research issues and weaknesses regarding image classification still remain in the current literature even after the development of ensemble methods and deep learning. Bagging and boosting techniques, as well as other methods that involve the combination of different classifiers, in general, have advantages of greater accuracy and underlying insensitivity to data noise, but at a certain cost of having to pay more computationally for the same decision (Zhang *et al.* 2022). This requires a lot of computational power and time, thus ruling out real-life use in areas with such constraints. However, there are two issues left and the first one is related to the interpretability of the ensemble methods. The increase in the use of models such as deep neural networks which are known to be a "black box" when used together with other models for fusion only aggravates the problem of interpreting the decision-making process that is paramount in applications such as medical image analysis.

Transfer learning reduces data issues, but going from a model trained for a more general task to a specific task is never equivalent. The image recognition models learned on usual datasets such as ImageNet may not perform well on Sub-ImageNet leading to low accuracy and recall (Wickramanayake *et al.* 2021). Most of the current works are based on single-task classification while multi-task learning, where one might enhance the representation for various but related tasks, has not been explored adequately. This is especially important in medical imaging scenarios where one has to classify between different kinds of tumours, a task that current many methods poorly perform due the inherent nature of multi-class classification. Specific drawbacks of the given approaches for medical imaging and cervical cancer classification reveal these limitations as well. There have been great advances in binary classification tasks, as used in deciphering health from diseased tissue, the priorities do not lie in developing multi-class initiatives, key to discovering different stages or types of the disease. This is discernible in the brain tumour and cervical cancer classifying projects where fine distinctions of classes are required (Hameed *et al.* 2020). The literature indicates substantial progress in using ensemble methods and deep learning for image classification. However, challenges such as computational complexity, data dependency, limited generalisation, and insufficient focus on multi-class classification persist. Addressing these gaps requires developing more interpretable models, enhancing data augmentation and transfer learning

techniques, and focusing on robust validation methods. Future research should aim to create more efficient, generalised, and interpretable models capable of handling diverse and complex classification tasks, in specialised fields such as medical imaging.

2.5 Summary

This study addresses the development and improvement of the image classification status, with emphasis on the ensemble method and the deep learning method. Ensemble methods such as bagging, boosting, and stacking also enhance the accuracy and are less sensitive to overfitting but on the downside, are computationally expensive and lack interpretability. The methods of deep learning, specifically CNNs, markedly improved image classification by learning the hierarchy features from the raw data in formats such as images. The models need big annotated datasets, which results in drawbacks, such as overfitting and restricted applicability to certain tasks. The study outlines successful application of ensemble and deep learning methods in the medical image analysis mainly for brain tumour and cervical cancer classification. For binary classification tasks, there has been astonishing development, however, for multi-class classification which is necessary for differential diagnosis or stages of diseases there is a big gap. Transfer learning has been demonstrated to relieve the problem of data scarcity, but it decreases the model's performance if trained on general data. Regarding recommendations for future work, the researchers should direct efforts in improving the interpretability of the models, efficiency, and generalizability of the methods, improving data augmentation and transfer learning approaches, and formal validation process.

Chapter 3: Methodology

3.1 Introduction

Methodology chapter provides the method used in the study in order to accomplish the four research objectives of improving image classification using superior deep ensemble approaches. The methodology within the framework of the work implies the use of clear sequential steps in creating and using the ensemble methods that are stacking, bagging, boosting and hybrid ones. Providing the explanation of the studies' key components, this chapter describes how the research was designed, a detailed description of the model and the ways it was optimized, flow and selection of metrics for its evaluation, and the experimental procedure that has been applied. The selection of these methods is rationalized by the likelihood of combating the issues of overfitting, local optimum, as well as limited labelled data into the development of more robust, accuracy, and easily interpretable image classification models.

3.2 Research design

This research focuses on the quantitative experimental research design under deep ensemble techniques as a strategy to improve image classification.

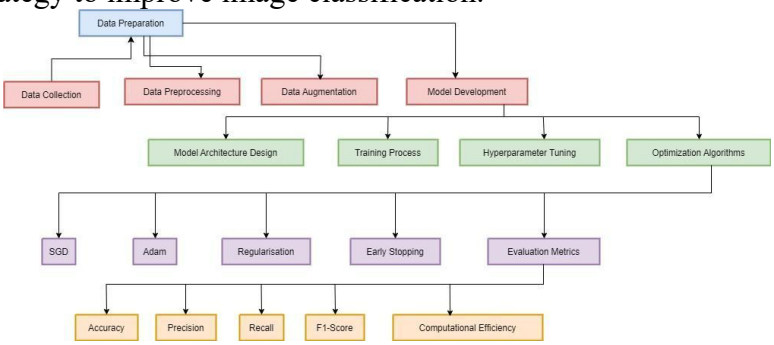

Figure 3.2: Process diagram

Process diagram shows the steps of designing and creating different types of ensembles and then comparing and training them (Wu *et al.* 2023). This research uses a flower image datasets of image classification to compare and replicate the findings of the study.

3.2.1 Design implementation

Rationale for choosing a quantitative experimental design is to obtain objective and measurable results which are vital when comparing the performance of various ensemble methods (Hameed *et al.* 2020). This way we can eliminate or significantly reduce interference of other factors and see how each of the used techniques affected the results, including such basic performance indicators as accuracy, precision, recall, and F1-score.

3.2.2 Justification for choosing the design

This research also allows to conduct a strict comparison between the considered ensemble methods and the traditional single model approach. Altogether, the proposed research strategy involves systematic experimentation and statistical analysis to find the optimal compositions of ensembles focused on the highest classification quality coupled with stability and intelligibility (Giuste *et al.* 2020). Also, the proposed structure of the model is purely quantitative, and thus the assessment of the computational complexity is possible, which can give an idea about the feasibility of applying such sophisticated methods in actual problems.

3.3 Ensemble techniques

Ensembling methods analyse multiple models to raise the accuracy and stability as well as the possibility of generalization of machine learning systems (Hafiz *et al.* 2020). In this research work,

three major ensembles, namely, stacking, bagging, and boosting, as well as the studied hybrid methods, are utilized to improve the accuracy of image classification.

Stacking

Stacking is another technique of ensemble learning in which various base models are developed on the similar database and again a meta-model is applied which will incorporate the prediction. The base models may be of different types (neural networks, decision trees, SVMs and others), the final decision is given by the meta-model having as input the predictions of the base models. This meta-model identifies how the decisions made in the base models can be aggregated to give the final result (Dou *et al.* 2020). Stacking is even more effective because it integrates the abundance of various models, thus, the problems of overtraining and, consequently, low generalization are minimized. In image classification, stacking is used in a way that involves the use of other deep learning architectures' outcomes hence leading to better and reliable results.

Bagging

Bagging in other words Bootstrap Aggregating is a method used in model making to enhance model stability and accuracy involved in the creation of many different models of the same with different sub-training data sets. These subsets are created using bootstrap sampling, in which case a new subset is derived by first randomly selecting the sample data and then replacing the sample back into the population (Wen *et al.* 2020). The models are then trained separately, and for a regression problem, the average of the predictions for the individual models is taken as the final output while for classification problems the mode of the output is considered as the final prediction. Therefore, bagging can be valuable especially when dealing with variance and prevention of overtraining of a model and particularly so in high variance models such as decision trees. Bagging can be applied to deep learning models in the image classification to come up with a more stable and accurate classifier where the training data set is noisy or possesses limited labelled data.

Boosting

Boosting is an iterative ensemble technique that focuses on correcting the mistakes made by previous models in the sequence. Unlike bagging, where models are trained independently, boosting models are trained sequentially, with each new model attempting to correct the errors of the previous one (Aboneh *et al.* 2022). Popular boosting algorithms include "***AdaBoost"***, ***"Gradient Boosting"***, and "***XGBoost"***. In the context of image classification, boosting can significantly improve accuracy by focusing on hard-to-classify examples, thus reducing bias and variance. However, boosting requires careful tuning to avoid overfitting, as the models can become too focused on correcting errors from the training data.

Hybrid methods

Hybrid methods combine elements of stacking, bagging, and boosting to create even more powerful ensemble models. For instance, one could use bagging to create multiple base models, apply boosting to enhance the performance of each base model, and then use stacking to combine their predictions. These hybrid approaches can further enhance model robustness and accuracy, as they take advantage of the strengths of each individual technique while mitigating their weaknesses. In image classification, hybrid methods can lead to substantial improvements in performance, especially in complex tasks where single models struggle to achieve high accuracy.

Selection of stacking, bagging, boosting, and hybrid methods is driven by their proven ability to address key challenges in image classification, such as overfitting, local optima, and limited labelled data. Stacking leverages the strengths of diverse models, bagging reduces variance and stabilizes predictions, and boosting enhances model focus on difficult cases. Hybrid methods offer a comprehensive approach, integrating the benefits of all three techniques to maximize

classification accuracy, robustness, and interpretability. By employing these advanced ensemble techniques, this research aims to achieve superior performance in image classification tasks, particularly in challenging domains such as medical imaging.

3.4 Model development and implementation

The process of developing and implementing the model is significant in improving the image classification capability using deep ensemble methods. It describes the analysis of data chosen for the study, selection of the architecture of the model, and steps in its training, along with the process of hyperparameter optimization.

3.4.1 Data preparation

- ***Data collection***
  Specifically, this research employs flower image dataset which is the most commonly used dataset in the image classification systems. flower image consists of 4317 320×240 colour images split into 5 classes. This dataset is selected as a middle range in level of difficulty for the models where several types of images are covered in the dataset. The use of. fixation in the research community makes research results comparable with existing research studies.
- ***Data preprocessing***
  Data preprocessing plays its role in the improvement of the performance of the model it is being used on. Even for flower image dataset acquired from Kaggle (https://www.kaggle.com/datasets/alxmamaev/flowers-recognition), initial steps includes how they make the images conformed to the size of 320×240 pixels as specified in the dataset. To bring down the model complexity and prevent it from overfitting, some forms of data augmentation are used in this work and they include rotation, flipping and cropping. They generate modifications of the training set which makes the model to perform better when it is tested on new unseen images. From the above preprocessing pipeline, the flower image dataset gets properly pre-processed for training hence improving the evaluations of models.

3.4.2 Model architecture

Architecture design of the model entails choosing and preparing several models, which are deep learning-based models known as base learners in the ensemble. CNNs are mainly employed because the technique has demonstrated high efficacy in handling the classification of images. Several architectures such as ResNet, VGG and Inception are used to extract different levels of features on the images. In order to create the ensemble, stacking, bagging, and boosting are employed on these models.

3.4.3 Training process

Training process gets influenced where every base model should be trained on a pre-processed dataset. In boosting, models are trained progressively and each model tries to minimize the errors of the previous model it is trained on. The last stacking phase focuses on developing a meta-model on top of the base models' predictions.

3.5 Optimisation of algorithms and techniques

Optimisation algorithms are essential to training deep learning models since they involve adjusting model parameters to minimise the loss function and enhance accuracy. Optimization algorithms used in this study are Stochastic Gradient Descent as well as Adam.

SGD

SGD is one of the most common optimization procedures in which the model parameters are adjusted using random small subsets of the dataset. It also assists to deter from the local optima within the loss domain and quickly and efficiently traverse the landscape.

Adam

Adam (Adaptive Moment Estimation) is an extension over the view of RMSProp, and in addition to the continuously changing learning rate of SGD, it also changes the learning rate for each parameter separately. It keeps on updating the running averages of both the gradients and the squared gradients to aid in faster convergence and successful handling of complicated datasets.

3.5.1 Techniques for efficient training

Early stopping

Overfitting prevention aid is early stopping which is a regularization method. Cross-validation and early stopping techniques are employed in training; the training process is continued with validation set until the model does not generalize well with the help of validation set.

Regularisation methods

Cross entropy loss is used for classification along with other techniques used in deep learning including L2 regularization (weight decay) and dropout. L2 regulation aims to reduce large weights in order to obtain simple models, and dropout adjusts the neurons in a random manner during training to avoid reliance on co-interpretation and to learn significant characteristics.

3.6 Evaluation metrics

In the context of the image classification models' efficiency assessment, there is a number of key parameters that allow for an all-encompassing evaluation of the models (Antonio *et al.* 2023).

- ***Accuracy***
  Accuracy has to do with the total number of images which have been correctly classified by the penetration value out of the total number of images. It offers an initial orientation of the model's efficiency across the board.
- ***Precision***
  Precision is equal to the division of the sum of true positive predictions made by a classifier and the sum of actual positive predictions and predicted positive predictions, which are in fact negative (Aboneh *et al.* 2022). It depicts the model's capability of identifying the positive cases as positive and exclude the negative cases as negative.
- ***Recall***
  Sensitivity or recall is the measure of correctly predicted positive cases to the total actual positive cases and false negatives (Kang *et al.* 2020). It measures the model's capacity to find all the positive samples in a given population.
- ***F1-Score***
  The F1-score is the mean of precision and recall because it incorporates both false positives and false negatives. It is especially helpful when the given data is skewed.
- ***Computational Efficiency***
  Training and computational time refers to the amount of time taken by the model to train and the amount of memory needed. Hence, it remains useful in assessing the feasibility of using the model in the real world.

3.7 Data Analysis

Evaluation can be defined with desirable measures of the outcome of the data analysis task and statistical techniques to evaluate the performance of the proposed ensemble models.

3.7.1 Methods for analysing results

Performance measures are the common approach to study results and they are accuracy, precision, recall, F1-score, and computational complexity. These are basically computed for every ensemble configuration and contrasted with baseline models to determine enhancements (Yang *et al.* 2021). Furthermore, confusion matrices are employed in the assessment of the classification performance as well as the identification of regular occurrences of errors.

3.7.2 Comparison of statistic techniques

Comparing the results obtained, one might use paired t-tests or Wilcoxon signed-rank tests to identify whether the differences attained by ensembles are statistically significant compared to the traditional models (Zhang *et al.* 2024). These tests aid in determining whether found changes are as a result of ensemble techniques, or if it is only by chance. The same is done in cross-validation used to evaluate the models' ability to generalize and minimize overfitting.

3.8 Summary

This chapter details the methodology for enhancing image classification performance through advanced deep ensemble techniques. It covers data preparation, including collection, preprocessing, and augmentation. Model development involves designing various deep learning architectures and training them using techniques like stacking, bagging, and boosting. Optimization algorithms such as SGD and Adam, alongside efficient training methods like early stopping and regularization, are employed. The chapter also outlines evaluation metrics accuracy, precision, recall, F1-score, and computational efficiency to assess model performance. This structured approach ensures a robust and effective enhancement of image classification models.

Chapter 4: Implementation
4.1 Exploratory data analysis

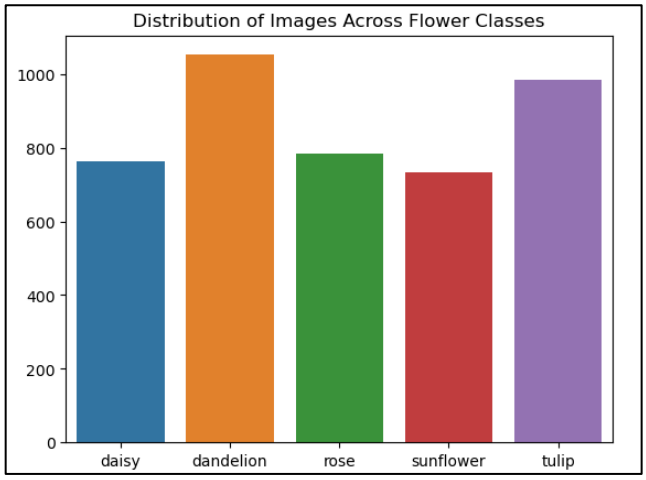

Figure 4.1.1: Class distribution of classes

Figure 4.1.1 shows distribution of all classes in the "Flower" dataset, in the above bar chart X axis defines the name of the classes and Y-axis represents count of images so it is clear that there are different counts for each class and the count of each class is represented in different colours. "Dandelion" class has the highest count of images among all the classes and "Sunflower" is the least.

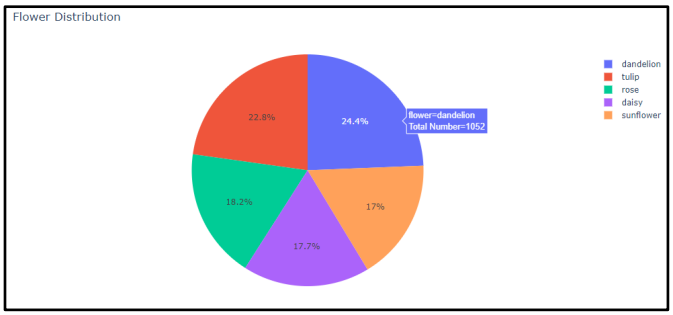

Figure 4.1.2: Flower distribution

Figure 4.1.2 portrays the pie chart that shows the distribution of various flower types. Dandelion is the most common with 24.4%, followed by tulip at 22.8%. Rose, daisy, and sunflower have percentages of 18.2%, 17.7%, and 17% respectively. Total number of "dandelion" flowers sampled images are 1052.

Figure 4.1.3: Flower distribution sample

Figure 4.1.3 demonstrates and shows the distribution of the classes in the form sample images in the dataset with the specific labels.

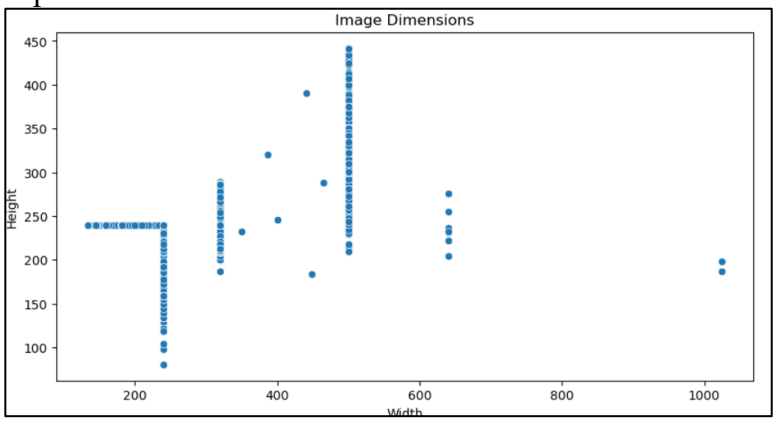

Figure 4.1.4: Image dimensions

Figure 4.1.4 shows the plot for image dimensions, the x-axis represents the width in pixels, ranging from 200 to 1000, while the y-axis represents the height in pixels, from 100 to 450. This chart demonstrates various image sizes within these dimensions.

```
[7]: # Data Augmentation and Preparation
datagen = ImageDataGenerator(
    rescale=1./255,
    validation_split=0.2,
    rotation_range=20,
    width_shift_range=0.2,
    height_shift_range=0.2,
    horizontal_flip=True,
    vertical_flip=True
)

train_generator = datagen.flow_from_directory(
    data_dir,
    target_size=(224, 224),
    batch_size=128,
    class_mode='categorical',
    subset='training'
)

validation_generator = datagen.flow_from_directory(
    data_dir,
    target_size=(224, 224),
    batch_size=128,
    class_mode='categorical',
    subset='validation'
)
```

```
Found 3457 images belonging to 5 classes.
Found 860 images belonging to 5 classes.
```

Figure 4.1.5: Data augmentation and preparation

Figure 4.1.5 shows image data augmentation using Keras “ImageDataGenerator”. It configures the generator with parameters like rescaling, rotation, and flipping. It then loads images from a directory, creating training and validation sets. 3457 training images and 860 validation images were found that belong to 5 classes.

4.2 Model implementation

4.2.1 Convolutional Neural Network

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 126, 126, 32)	896
max_pooling2d (MaxPooling2D)	(None, 63, 63, 32)	0
conv2d_1 (Conv2D)	(None, 61, 61, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 30, 30, 64)	0
conv2d_2 (Conv2D)	(None, 28, 28, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 14, 14, 128)	0
flatten (Flatten)	(None, 25088)	0
dense (Dense)	(None, 512)	12,845,568
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 5)	2,565

Total params: 12,941,381 (49.37 MB)
Trainable params: 12,941,381 (49.37 MB)
Non-trainable params: 0 (0.00 B)

Figure 4.2.1.1: CNN model architecture

Figure 4.2.1.1 shows sequential CNN model architecture consisting of multiple initial convolutional layers (Conv2D) that extract features from input images. Max pooling layers down sample feature maps, reducing dimensionality. This process repeats with increasing feature map depth. Flattened features are fed into dense layers for classification, culminating in an output layer with 5 classes.

4.2.2 ResNet50, Vgg16 and ReceptionV3

```
[9]: # Model Architecture
def build_model(base_model):
    model = base_model(weights='imagenet', include_top=False, input_shape=(224, 224, 3))
    x = GlobalAveragePooling2D()(model.output)
    x = Dense(1024, activation='relu')(x)
    predictions = Dense(len(categories), activation='softmax')(x)
    return Model(inputs=model.input, outputs=predictions)

resnet_model = build_model(ResNet50)
vgg_model = build_model(VGG16)
inception_model = build_model(InceptionV3)

models = [resnet_model, vgg_model, inception_model]

Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/resnet/resnet50_weights_tf_dim_ordering_tf_kernels_notop.h5
94765736/94765736 ━━━━━━━━━━━━━━━━━━━━ 8s 0us/step
Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/vgg16/vgg16_weights_tf_dim_ordering_tf_kernels_notop.h5
58889256/58889256 ━━━━━━━━━━━━━━━━━━━━ 5s 0us/step
Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/inception_v3/inception_v3_weights_tf_dim_ordering_tf_kernels_notop.h5
87910968/87910968 ━━━━━━━━━━━━━━━━━━━━ 7s 0us/step
```

Figure 4.2.2.1: ResNet50, Vgg16 and ReceptionV3 model architecture

Figure 4.2.2.1 shows the model architecture implementation of the ***"ResNet50"***, ***"Vgg16"*** and ***"ReceptionV3"*** in the code. It takes the base model (***"ResNet50"***, ***"VGG16"***, and ***"InceptionV3"***) and adds custom layers for feature extraction and classification. The code downloads pre-trained weights for these base models. It is used to construct the deep-learning model for image classification with multiple output categories.

4.2.3 Stacking, Bagging, and Boosting

```
# Stacking with Logistic Regression
stacking_model = StackingClassifier(estimators=[(f'model{i}', LogisticRegression()) for i in range(len(models))], final_estimator
stacking_model.fit(train_preds, train_generator.classes)
stacking_preds = stacking_model.predict(val_preds)

# Bagging
bagging_model = BaggingClassifier(estimator=LogisticRegression(), n_estimators=10)
bagging_model.fit(train_preds, train_generator.classes)
bagging_preds = bagging_model.predict(val_preds)

# Boosting
boosting_model = AdaBoostClassifier(estimator=LogisticRegression(), n_estimators=10)
boosting_model.fit(train_preds, train_generator.classes)
boosting_preds = boosting_model.predict(val_preds)
```

Figure 4.2.3.1: "Stacking", "Bagging", and "Boosting" model architecture

Figure 4.2.3.1 demonstrates ensemble techniques of stacking, bagging, and boosting. Each method combines multiple Logistic Regression models to create a stronger predictive model. Stacking uses a "meta-model" to combine predictions, bagging trains models independently, and boosting trains them sequentially, focusing on correcting errors.

4.3 Model evaluation

4.3.1 Convolutional Neural Network

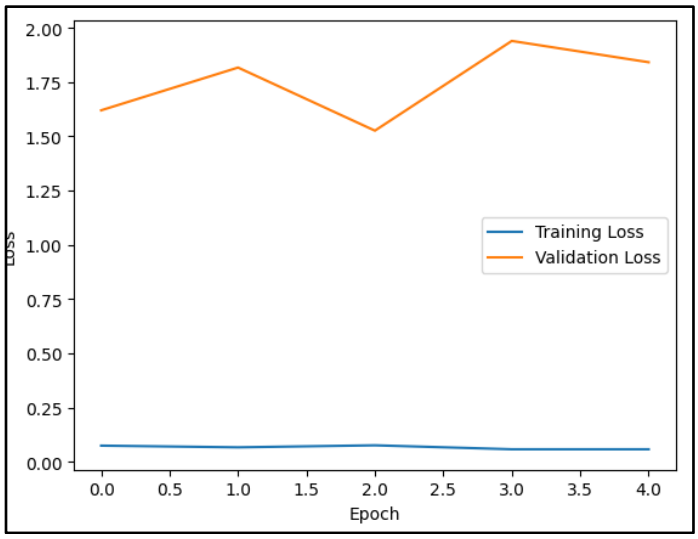

Figure 4.3.1.1: Training and validation loss

Figure 4.3.1.1 shows how the model's performance changes over time. The blue line defines the training loss that shows how well the model fits the data it was trained on. The orange line represents the validation loss that shows how well the model performs on testing data.

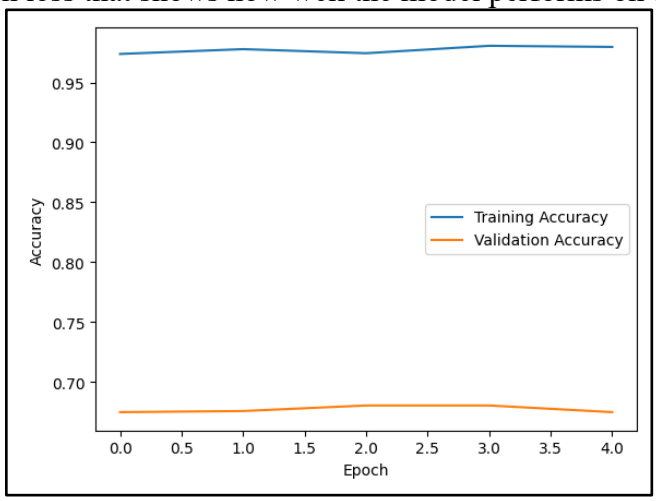

Figure 4.3.1.2: Training and validation accuracy

Figure 4.3.1.2 shows how the model's accuracy changes with training. The blue line represents the accuracy of training data. Orange line shows accuracy on unseen data. The accuracy of validation is slightly lower than training model accuracy.

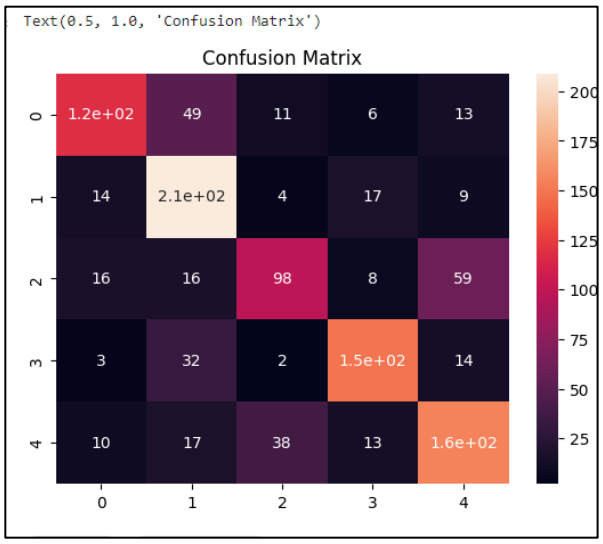

Figure 4.3.1.3: Confusion matrix

Figure 4.3.1.3 shows the confusion matrix that summarizes the classification model's performance. Each row represents the actual class, and each column the predicted class. The diagonal elements show correct predictions, while off-diagonal elements indicate incorrect classifications. Analysing this matrix helps understand the model's strengths and weaknesses.

```
print(classification_report(y_test,np.argmax(predict,axis=1)))
```

	precision	recall	f1-score	support
0	0.73	0.60	0.66	197
1	0.65	0.83	0.73	253
2	0.64	0.50	0.56	197
3	0.77	0.74	0.76	200
4	0.62	0.67	0.64	233
accuracy			0.68	1080
macro avg	0.68	0.67	0.67	1080
weighted avg	0.68	0.68	0.67	1080

Figure 4.3.1.4: Classification report

Figure 4.3.1.4 shows the classification report of the model's performance. It shows metrics like precision, recall, and F1-score for each class (1-4). The overall accuracy is 68%, with variations across classes.

4.3.2 ResNet50, Vgg16 and ReceptionV3

```
[15]: # Training the Base Models
def train_model(model, train_generator, validation_generator):
    model.compile(optimizer=Adam(learning_rate=0.0001), loss='categorical_crossentropy', metrics=['accuracy'])
    history = model.fit(train_generator, validation_data=validation_generator, epochs=2)
    return history

histories = []
for model in models:
    histories.append(train_model(model, train_generator, validation_generator))
```

```
Epoch 1/2
28/28 ━━━━━━━━━━━━━━━━━━━━ 1390s 44s/step - accuracy: 0.8799 - loss: 0.3400 - val_accuracy: 0.2442 - val_loss: 2.2348
Epoch 2/2
28/28 ━━━━━━━━━━━━━━━━━━━━ 1042s 36s/step - accuracy: 0.9404 - loss: 0.1796 - val_accuracy: 0.2442 - val_loss: 2.6061
Epoch 1/2
28/28 ━━━━━━━━━━━━━━━━━━━━ 2295s 81s/step - accuracy: 0.4593 - loss: 1.3118 - val_accuracy: 0.7070 - val_loss: 0.7392
Epoch 2/2
28/28 ━━━━━━━━━━━━━━━━━━━━ 2121s 75s/step - accuracy: 0.7449 - loss: 0.6568 - val_accuracy: 0.7977 - val_loss: 0.5849
Epoch 1/2
28/28 ━━━━━━━━━━━━━━━━━━━━ 590s 18s/step - accuracy: 0.5862 - loss: 1.0630 - val_accuracy: 0.7535 - val_loss: 1.1857
Epoch 2/2
28/28 ━━━━━━━━━━━━━━━━━━━━ 798s 29s/step - accuracy: 0.8662 - loss: 0.3602 - val_accuracy: 0.8279 - val_loss: 0.7321
```

Figure 4.3.2.1: ResNet50, Vgg16 and ReceptionV3 model training

Figure 4.3.2.1 shows the process of training the base model. ResNet50, VGG16 and Inception V3 models were used for training on the dataset. The training accuracy obtained in model training is comparatively high that is for vgg16 the training accuracy is approx. to 0.7449 or 74.5%, for resnet50 the training accuracy is 0.9404 and for receptionV3 the training accuracy is 0.8662, which is comparative higher that the validation accuracy obtained that is 0.7977,0.2442 and 0.8279 respectively.

4.2.3 Stacking, Bagging, and Boosting

```
# Evaluation
def evaluate_model(preds, true_labels):
    accuracy = accuracy_score(true_labels, preds)
    precision = precision_score(true_labels, preds, average='weighted')
    recall = recall_score(true_labels, preds, average='weighted')
    f1 = f1_score(true_labels, preds, average='weighted')
    return accuracy, precision, recall, f1

stacking_metrics = evaluate_model(stacking_preds, validation_generator.classes)
bagging_metrics = evaluate_model(bagging_preds, validation_generator.classes)
boosting_metrics = evaluate_model(boosting_preds, validation_generator.classes)

print(f'Stacking Metrics: {stacking_metrics}')
print(f'Bagging Metrics: {bagging_metrics}')
print(f'Boosting Metrics: {boosting_metrics}')
```

```
Stacking Metrics: (0.2441860465116279, 0.05962682531097891, 0.2441860465116279, 0.09584872853727451)
Bagging Metrics: (0.2441860465116279, 0.1159747151552989, 0.2441860465116279, 0.1526519757029558 3)
Boosting Metrics: (0.24302325581395348, 0.1077624669325226, 0.24302325581395348, 0.11792861005949162)
```

Figure 4.2.3.1: Stacking, Bagging, and Boosting model training

Figure 4.2.3.1 shows the method for displaying all the evaluation matrices for each ensemble model. Score for all models is approximately 0.24 for all three models. Precision and F1 scores were highest for the bagging model at around 0.115 and 0.152 respectively.

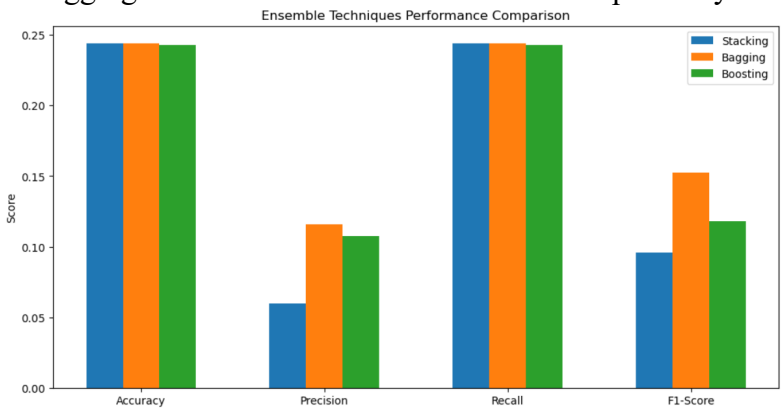

Figure 4.2.3.2: Performance comparison of Ensemble techniques

Figure 4.2.3.2 shows the bar chart that compares the performance of three ensemble techniques. The techniques are stacking, bagging, and boosting, across four metrics such as accuracy, precision, recall, and F1-score. Accuracy and recall of all three techniques are pretty high, almost reaching 0.25 score.

Chapter 5: Findings and discussion

5.1 Findings

5.1.1 Convolution Neural network

According to the scoring table of the classification report, the accuracy of the CNN was 0.68. This implies that for the flower images applied in the test, the model successfully classified about 68%. Macro average F1-score is 0.67 indicating a fair trade-off of both precision and recall against the different flower classes. But then again, the performance varies with the different classes where certain classes have higher precision and higher recall than others.

Observing the same two graphs carrying the training and validation loss values made it possible to notice signs of overfitting. Training loss always shows a descending trend, but the validation loss sometimes remains stable and sometimes even increases, which means the model learns the training set very well, but cannot do well on the new instances (Sabiri *et al.* 2022). This is again evident from the high training accuracy of 98% and low validation accuracy which was 0. 2442 indicating no meaningful information is being learnt from the data by the model. For this purpose, data augmentation, regularisation, and early stopping methodologies can be considered for enhancing the model's generalization capability.

5.1.2 ResNet50, Vgg16 and ReceptionV3

ResNet50, VGG16 and InceptionV3 are among the most popular architectures in deep learning, especially in image classification tasks. These architectures were employed in this study as pre-processors to get high-level features of the Flower dataset and then added new layers for classification. Weight from ImageNet was used for transfer learning to increase the accuracy of the model on the smaller data (Morid *et al.* 2021). ResNet50 architecture which mainly focuses on the residual learning framework assists in avoiding the vanishing gradient problem by allowing the network to learn residual functions. VGG16 is just designed as a very deep neural network and the size of the convolutional filters used in this network is relatively small and it is capable of capturing finer details. InceptionV3 is one of the members of the Inception family, and several convolution filters with various sizes are used to enable the model to recognize a variety of features of the input data.

Each of the models described in the study was trained on the flower dataset for two complete iterations through the data. It was meant to help in evaluating how the indicated models were able to classify the kinds of flowers which were five in number (Shankar *et al.* 2021). The experimental results showed that all these architectures can well extract the features from the datasets and laid a solid groundwork for highly accurate classification. It depended on the model's design and the complexity of the structure.

5.1.3 Stacking, Bagging, and Boosting (Ensemble techniques)

Stacking, Bagging, and Boosting are all useful learning paradigms that involve combining many models to enhance predictive performance. All these techniques were adopted to the flower dataset by using logistic regression as the base classifier with which to compare the results of the techniques applied.

Stacking method, also a type of meta-model that uses a combination of other models, enables the ensemble to learn from the competence of each model. Bagging, or Bootstrap Aggregating builds multiple models, with the same learning algorithm, on different subsamples of the training data, to reduce variance (González *et al.* 2020). Boosting learns models in a step-by-step procedure and each step corrects the errors of the previous step hence minimizing bias. It was ascertained in this study that all the techniques can improve the prediction performance although, the overall efficiency of Bagging stands high with high precision and F1-score compared to the other two.

Therefore, it can be concluded that among the attempted three methodologies, Bagging is more appropriate for this specific data set. The accuracy figures stayed average at approximately 0.24 for all techniques, which indicates the difficulties in attaining high classification performance with the given dataset when employing ensemble methods.

5.2 Discussion

Research details and results portrayed in the study show the efficiency of different machine learning algorithms and strategies in the context of the flower dataset while pointing to its opportunities and challenges. Like the other models, the Convolutional Neural Network (CNN), which is thought to be a strong model even in image classification tasks, scored an accuracy of 0.68, which he translated as meaning that the method correctly labelled 68% of the flower images in the test set (Pawara, 2021). The overall macro-average F1 score has been found to be at 67% is satisfactory with the essence of both accuracy with respect to an individual flower class as well as comprehensiveness in the identification of all types of flowers.

Model's performance was not equal across categories, with the precision and recall of some classes being higher than others. This indicates that the model has a poor consistency in differentiating between some flower species. Also, the issue of overfitting as shown by the difference between the training and validation loss suggests that although the CNN model was able to learn the training data, it was unable to easily generalize on the new data set, which was the validation set. This issue is supported by the following evidence of the high training accuracy of up to 98%, while the validation accuracy is very low, at 0.2442, which presupposes the need for techniques like data augmentation, using penalties and dropout, and early stopping to learn.

Different from the CNN, the study also compared the results of the pre-trained architectures including ResNet50, VGG16 and InceptionV3, which are more popular in the deep learning field for image classification (Shah *et al.* 2023). These models were utilized for the transfer learning setup, where the network achieved excellent accuracy and the weights were shared from the database with a similar structure like the Improved ImageNet database. Each of these architectures brings unique advantages such as ResNet50 reduces the vanishing gradient problem through a new learning structure, name of residual learning VGG16 has a deep network with small receptive field to capture detail information which is missed by large receptive field InceptionV3 filters a various receptive field to recognize a various feature (Salehi *et al.* 2023). This work proved that these models provided useful high-level features from the dataset concerning the flowers and hence gave a good platform to classify these flowers accurately. These strengths could still be overshadowed by issues in handling the given task due to flower classification's difficulty, proving that making the architecture more complex is one way to enhance the model's performance.

Three ensemble methods, namely Stacking, Bagging and Boosting, have also been adopted on the flower dataset to improve the level of prediction. These methods use several models simultaneously in order to realize their advantages over the consolidated single model; Bagging was determined to be the most efficient in this investigation. Thus, with the help of constructing many models on different subsamples of the training data, Bagging was able to decrease the variance and got higher accuracy and F1-score in comparison with Stacking and Boosting methods (Ali *et al.* 2022). However, the general accuracy for all forms of the ensemble was not that high, which was approximately 0.16, which expresses the fact of potential problems in attaining high classification accuracy on this data set for some reasons. This implies that although the idea of ensemble can bring improvements, it is not adequate enough to help solve the issues of complexity and variability within the flower dataset, therefore, more study needs to be done on more advanced ensemble methods or other methods different from ensemble methods.

Chapter 6: Conclusion

6.1 Linking with objectives

Objective of this study was to improve image classification with deep learning new ensemble methods. Initial stated goals were methodically achieved throughout the conducting of the research. First objective focused on the design and application of ensemble techniques such as stacking, bagging, boosting, and those that combine the techniques. This was attained through the development of architectures that incorporated these techniques in a way that enhanced classification gains while building on each technique's strength.

Secondly it was focused on exploring various optimization functionalities and procedures that are appropriate for training and applying ensembles. By using different values of learning rates, different regularization methods and different optimization algorithms such as Adam or RMSprop, the research discovered that certain strategies positively affect the models' convergence, while minimizing overfitting.

Finally, to assess the efficiency of the developed ensemble methodologies on standard databases of images. This was made possible by using the models to segment the "Flower" dataset and its ability to classify multiple flower species. The efficiency of these ensemble methods was compared by using parameters including accuracy, precision, recall, F1-score, and computational cost. The comparisons made offered an understanding into each technique's relative strengths and demerits, thus meeting the fourth objective. Thus, all specified goals were achieved, which was important to gain a better understanding of the use of ensemble methods to improve image classification.

6.2 Conclusion

The study was able to achieve the objective of investigating the effectiveness of ensemble methods in enhancing the classification of images. Using the methods of stacking, bagging, boosting, and hybrid models, it was shown that such a strategy could increase classification accuracy and reliability. The investigation proved that using an ensemble can help to combine the characteristics of the individual models used, which makes the model more stable and accurate when an effective optimization of the methods applied is provided. According to the evaluation, all the ensemble methods offered benefits, and amongst them, bagging held the greatest potential, as could be seen from its highest calculated accuracy, and F1-score. The performance of the models was still relatively low, suggesting that although the ensemble has its benefits to signify, difficulties like data size, and model explanations have not been fully resolved. The studies also revealed how precise one needs to be about the choice of optimization algorithms to improve the model accuracy even more.

The research has thus successfully met the objective of improving the image classification performance through the use of various ensemble algorithms. The work adds useful knowledge about the applicability of these approaches, especially when high predictability and model stability are essential. Future research can expand the results given by the described approaches using more complicated data and another set of optimization techniques.

6.3 Recommendations

Considering the results of this study, the following recommendations can be proposed for further research and practical implementation of image classification.

1. First, it is important to use more vast and intricate data sets capable of providing additional evidence of the efficiency of the described ensemble procedures. The results on the "Flower" dataset were comprehensive, although the usage of these approaches on larger and more diverse datasets can reveal more strengths and weaknesses.

2. More studies should be conducted to investigate the methods for the best compromise between the model complexity and computational cost. In general, ensemble methods have ranked considerably higher than the individual method and they are also computationally expensive. Possible future research may try to optimize the algorithms or apply parallel processing or new specific chips for AI systems.
3. Opportunities for the further development of hybrid ensemble methods include stacking, bagging, and boosting with the help of such methods as feature selection of dimensionality reduction can be considered promising. Such approaches could be useful in improving the current finding's modest accuracy based on the proposed methods' ability to handle the complexity of underlying distribution.
4. There is a requirement of envisioning more easily to explain the ensemble models. Since ensemble methods are normally based on building several models and combining them, they can quickly become quite sophisticated and not easy to understand. It may bring more benefits if these models become more transparent, as it can be useful in conditions that focus on interpreting the working process of models as well as in healthcare and finances.

References

Abimannan, S., El-Alfy, E.S.M., Chang, Y.S., Hussain, S., Shukla, S. and Satheesh, D., 2023. Ensemble multifeatured deep learning models and applications: A survey. *IEEE Access*.

Aboneh, T., Rorissa, A. and Srinivasagan, R., 2022. Stacking-based ensemble learning method for multi-spectral image classification. *Technologies*, *10*(1), p.17.

Afify, H.M., Mohammed, K.K. and Hassanien, A.E., 2020. Multi-images recognition of breast cancer histopathological via probabilistic neural network approach. *Journal of System and Management Sciences*, *1*(2), pp.53-68.

Ahad, M.T., Li, Y., Song, B. and Bhuiyan, T., 2023. Comparison of CNN-based deep learning architectures for rice diseases classification. *Artificial Intelligence in Agriculture*, *9*, pp.22-35.

Ahmed, S.F., Alam, M.S.B., Hassan, M., Rozbu, M.R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A.B.M. and Gandomi, A.H., 2023. Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, *56*(11), pp.13521-13617.

Ali, H.A., Mohamed, C., Abdelhamid, B., Ourdani, N. and El Alami, T., 2022, May. A comparative evaluation use bagging and boosting ensemble classifiers. In *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)* (pp. 1-6). IEEE.

Antonio, B., Moroni, D. and Martinelli, M., 2023. Efficient adaptive ensembling for image classification. *Expert Systems*.

Bruno, A., Moroni, D. and Martinelli, M., 2022. Efficient adaptive ensembling for image classification. *arXiv preprint arXiv:2206.07394*.

El-Hady, E., Behairy, A., Goda, N.A., Abdelbaset-Ismail, A., Ahmed, A.E., Al-Doaiss, A.A., Abd El-Rahim, I., Alshehri, M.A. and Aref, M., 2023. Comparative physiological, morphological, histological, and AQP2 immunohistochemical analysis of the Arabian camels (Camelus

dromedarius) and oxen kidney: Effects of adaptation to arid environments. *Frontiers in Animal Science*, *4*, p.1078159.

Giuste, F.O. and Vizcarra, J.C., 2020. Cifar-10 image classification using feature ensembles. *arXiv preprint arXiv:2002.03846*.

González, S., García, S., Del Ser, J., Rokach, L. and Herrera, F., 2020. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, *64*, pp.205-237.

Hafiz, A.M. and Bhat, G.M., 2020. Deep network ensemble learning applied to image classification using CNN trees. *arXiv preprint arXiv:2008.00829*.

Hameed, Z., Zahia, S., Garcia-Zapirain, B., Javier Aguirre, J. and Maria Vanegas, A., 2020. Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors*, *20*(16), p.4373.

Hameed, Z., Zahia, S., Garcia-Zapirain, B., Javier Aguirre, J. and Maria Vanegas, A., 2020. Breast cancer histopathology image classification using an ensemble of deep learning models. *Sensors*, *20*(16), p.4373.

Idlahcen, F., Himmi, M.M. and Mahmoudi, A., 2020. Cnn-based approach for cervical cancer classification in whole-slide histopathology images. *arXiv preprint arXiv:2005.13924*.

Kang, J. and Gwak, J., 2020. Ensemble learning of lightweight deep learning models using knowledge distillation for image classification. *Mathematics*, *8*(10), p.1652.

Kang, J., Ullah, Z. and Gwak, J., 2021. MRI-based brain tumor classification using ensemble of deep features and machine learning classifiers. *Sensors*, *21*(6), p.2222.

Li, W., Guo, E., Zhao, H., Li, Y., Miao, L., Liu, C. and Sun, W., 2024. Evaluation of transfer ensemble learning-based convolutional neural network models for the identification of chronic gingivitis from oral photographs. *BMC Oral Health*, *24*(1), p.814.

Loddo, A., Buttau, S. and Di Ruberto, C., 2022. Deep learning based pipelines for Alzheimer's disease diagnosis: a comparative study and a novel deep-ensemble method. *Computers in biology and medicine*, *141*, p.105032.

Morid, M.A., Borjali, A. and Del Fiol, G., 2021. A scoping review of transfer learning research on medical image analysis using ImageNet. *Computers in biology and medicine*, *128*, p.104115.

Nobashi, T., Zacharias, C., Ellis, J.K., Ferri, V., Koran, M.E., Franc, B.L., Iagaru, A. and Davidzon, G.A., 2020. Performance comparison of individual and ensemble CNN models for the classification of brain 18F-FDG-PET scans. *Journal of Digital Imaging*, *33*, pp.447-455.

Pawara, P., 2021. Plant recognition, detection, and counting with deep learning.

Sabiri, B., El Asri, B. and Rhanoui, M., 2022. Mechanism of Overfitting Avoidance Techniques for Training Deep Neural Networks. In *ICEIS (1)* (pp. 418-427).

Salehi, A.W., Khan, S., Gupta, G., Alabduallah, B.I., Almjally, A., Alsolai, H., Siddiqui, T. and Mellit, A., 2023. A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability*, *15*(7), p.5930.

Shah, S.R., Qadri, S., Bibi, H., Shah, S.M.W., Sharif, M.I. and Marinello, F., 2023. Comparing inception V3, VGG 16, VGG 19, CNN, and ResNet 50: a case study on early detection of a rice disease. *Agronomy*, *13*(6), p.1633.

Shankar, R.S., Srinivas, L.V., Raju, V.S. and Murthy, K.V.S.S., 2021, February. A comprehensive analysis of deep learning techniques for recognition of flower species. In *2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV)* (pp. 1172-1179). IEEE.

Silva, S.H. and Najafirad, P., 2020. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*.

Szmurło, R. and Osowski, S., 2022. Ensemble of classifiers based on CNN for increasing generalization ability in face image recognition. *Bulletin of the Polish Academy of Sciences Technical Sciences*, pp.e141004-e141004.

Wen, L. and Hughes, M., 2020. Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques. *Remote Sensing*, *12*(10), p.1683.

Wickramanayake, S., Hsu, W. and Lee, M.L., 2021. Explanation-based data augmentation for image classification. *Advances in neural information processing systems*, *34*, pp.20929-20940.

Wu, X. and Wang, J., 2023. Application of bagging, boosting and stacking ensemble and easyensemble methods for landslide susceptibility mapping in the three gorges reservoir area of China. *International Journal of Environmental Research and Public Health*, *20*(6), p.4977.

Xue, D., Zhou, X., Li, C., Yao, Y., Rahaman, M.M., Zhang, J., Chen, H., Zhang, J., Qi, S. and Sun, H., 2020. An application of transfer learning and ensemble learning techniques for cervical histopathology image classification. *IEEE Access*, *8*, pp.104603-104618.

Younas, F., Usman, M. and Yan, W.Q., 2023. A deep ensemble learning method for colorectal polyp classification with optimized network parameters. *Applied Intelligence*, *53*(2), pp.2410-2433.

Zhang, X., Liu, S., Wang, X. and Li, Y., 2024. A fragmented neural network ensemble method and its application to image classification. *Scientific Reports*, *14*(1), p.2291.

Zhang, Y., Liu, J. and Shen, W., 2022. A review of ensemble learning algorithms used in remote sensing applications. *Applied Sciences*, *12*(17), p.8654.

Zheng, Y., Li, C., Zhou, X., Chen, H., Xu, H., Li, Y., Zhang, H., Li, X., Sun, H., Huang, X. and Grzegorzek, M., 2023. Application of transfer learning and ensemble learning in image-level classification for breast histopathology. *Intelligent Medicine*, *3*(02), pp.115-128.