INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Stacked Deep Ensemble Model for Learning Disability Detection

K Soumya[1]*, P Joseph Charles[2]*, S Kavitha[3]

1 Research Scholar ,St.Joseph's College(Autonomous),Affiliated to Bharathidasan University,Tiruchirappalli-620 002,Tamil Nadu,India
2 St.Joseph's College(Autonomous),Affiliated to Bharathidasan University,Tiruchirappalli-620 002,Tamil Nadu,India
3 Department of Computer Science, Christ University, Bangalore

OPEN ACCESS

Received: 21/09/2025
Accepted: 21/10/2025
Published: 16/11/2025

Citation: Soumya K, Joseph Charles P, Kavitha S (2025) Stacked Deep Ensemble Model for Learning Disability Detection. Indian Journal of Science and Technology 18(42): 3327-3338. https://doi.org/10.17485/IJST/v18i42.1612

* **Corresponding authors**.

soumyak.research@gmail.com

josephcharles_cs1@mail.sjctni.edu

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: Many students face learning difficulties and recognizing them early is vital for timely support. This study set out to explore handwriting analysis as a way to identify such difficulties with greater accuracy. **Method:** Four well-known convolutional neural networks—AlexNet, VGG16, ResNet50, and InceptionV3—were tested on a dataset of handwritten samples. After comparing their baseline results, features from ResNet50 and InceptionV3 were combined and passed through a fully connected neural network to form a stacked ensemble model. **Findings:** ResNet50 gave the strongest single-model performance, with an accuracy of 93.6%, followed by InceptionV3 at 91.8%. VGG16 and AlexNet reached 88.4% and 86.7%. The ensemble improved on all of these, reaching 95.4% accuracy, with precision of 95.1%, recall of 94.7%, and an F1-score of 94.9%. **Novelty:** The study goes beyond using a single CNN by building a stacked deep ensemble that draws on the strengths of ResNet50 and InceptionV3. This combined approach improves reliability and accuracy, showing clear potential for practical use in early screening of learning disabilities from handwriting.

Keywords: Tranfer learning; CNN architecture; AlexNet; VGG16; ResNet50; Inception; Learning Disability

1 Introduction

Learning disabilities can be compared to hurdles in a race, making it difficult for some individuals to maintain the same learning pace as others, despite having equal intelligence and opportunities. These difficulties don't just affect grades, they can impact confidence, daily activities, and the support families need. Catching these problems early makes a real difference, allowing teachers and parents to guide children in ways that suit them best.

Handwriting offers a window into these challenges. By looking at how children write, we can find patterns that hint at learning difficulties. In this study, we explore several convolutional neural network models- AlexNet, ResNet50, Inception, and VGG16- to see how well they can identify such patterns. Our goal is to combine their strengths into

a single framework that could be useful both in classrooms and in clinics, giving faster insights than traditional tests while helping children get support sooner.

Several studies have explored automated ways to detect learning difficulties. For example, Prabhala et al.[1] developed a stacked ensemble framework combining MHA-LSTM, DBN, and TDNN models for detecting dysarthric speech disabilities, with a Genetic Algorithm optimizing the contribution of each base model. This method is effective in improving accuracy and robustness, achieving high detection performance and faster inference times. However, the approach relies solely on speech signal features, which may limit insights from other complementary clinical or behavioral indicators.

Handwriting analysis using deep learning has gained attention because it uses existing student work. Soumya et al.[2] tested CNN transfer learning models and found that ResNet50 performed best compared to AlexNet, VGG16, and Inception. Its deep architecture and skip connections allowed it to capture complex writing patterns. Even so, most studies focus on single models, which limit overall accuracy and robustness.

John D Mayfield et al.[3] suggested a new longitudinal analysis with a Video based Vision Transformer model, compared to standard recurrent neural networks, emphasizing the CNN-LSTM architecture and a Fusion Vision Transformer -LSTM model. Multisequence contrast-enhanced cervical spine MRIs from 703 patients, over data collected between 2002 and 2023, from two institutions were used. VGG16-LSTM outperformed ViViT in samples with two years of MRI data and had an AUC of 0.75 as compared to ViViT, which had 0.72. Nevertheless, both models had decreased precision when making exact EDSS classifications using classification and regression approaches.

Vajs et al.[4] looked at eye movement patterns to identify dyslexia. They studied factors like eye fixation time on each word, reading speed per word and how often the eyes move backward. Their models could classify dyslexic and nin-dyslexic readers and proving every small reading behaviour can be informative and useful for detection.

Kulasekara et al.[5] developed a game-based framework to categorize dyslexia, dysgraphia, and dyscalculia, and at the same time also offering personalized practice. Though these methods are innovative, they do not focus on handwriting features, which are crucial for early detection.

2 Methodology

A combined transfer learning approach is used to detect learning disabilities by studying handwriting patterns. The method utilizes convolutional neural network (CNN) pre-trained models—VGG16, InceptionV3, AlexNet, and ResNet50—to learn rich and hierarchical features from input images. These models, originally established on the ImageNet dataset, are trained on a specialized haandwritten image dataset categorized into three classes: correct, casual, and reverse handwriting patterns.

2.1 Data Acquisition and Preparation

In the present study, we used a dataset consisting of handwritten letter images categorized into three classes: *normal*, *reversed*, and *corrected*. This dataset combines uppercase letters from nist special database 19[6], lowercase letters from the kaggle dyslexia handwriting dataset[7], a publicly accessible dataset consisting of 50,000 images was employed, which were classified into three distinct categories: reversed, normal, and corrected. It is important to clarify that this classification applies specifically when the images constituting the hologram are reversed, thereby exhibiting the normal and corrected images concurrently. Prior to being input into the chosen CNN (convolutional neural network) architectures, collected images underwent preprocessing to implement the requisite modifications and adjustments. In the initial preprocessing phase of the dataset, each image was standardized to a fixed dimension of 30 X 30 pixels. This step is crucial, as it ensures that all classifiers receive input images of consistent size. By applying this resizing process, the input size was successfully stabilized, not only lowering the computational load but also enhancing the processing algorithms' efficiency[8]. Apart from the above-mentioned resizing operation, the intensities of the pixels in the image were normalized to a range of 0 to 1, as mentioned in Equation 4. Normalizing the pixel intensity values was important in standardizing pixel intensity values, hence improving the ability of the models to learn from the data and generalize based on the data. Considering the effects of combining information coming from different sources and the need to enhance the Richness and diversity of the training dataset, a series of data augmentation methods were utilized. The methods utilized included random rotation of figures, flipping both horizontally and vertically, as well as zooming. These augmentations enriched the training dataset by adding more variability, which is required for alleviating the risk of model overfitting and thus enhancing its performance on new data.

After preprocessing, the dataset was divided into training and validation sets to make model training and testing more effective. About 70% of the images (30,000) were used for training, while 20,000 images were set aside for validation. In machine learning, this type of split—often referred to as the 70-30 rule— with the majority of the data being used to train the model and a smaller portion remaining to assess its performance.

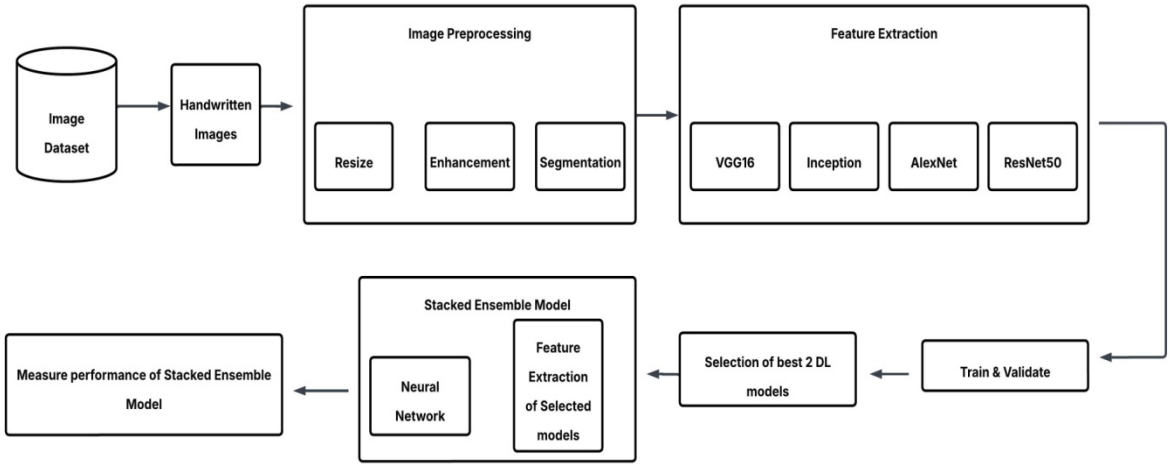

Fig 1. Flow Diagram of Proposed Model for Learning Disability Detection

The Four selected CNN models were trained using training dataset to find the in-depth features of the different classes such as corrected, normal, and reversed. The models were trained on augmented and normalized data to improve their understanding of classification tasks. To measure performance, the study used matrices like accuracy, precision, recall, and F1-score for checking how well the models could classify data. On the validation phase, CNN models were tested on new set of images that were not used during training.

The validation phase offered valuable insights into model behavior, identifying weaknesses and areas for improvement. A rigorous approach was followed, splitting the dataset into smaller, manageable training and testing subsets. Preprocessing and augmentation methods were utilized to enhance the efficiency and efficacy of CNN models for image classification.

2.2 Algorithm Selection and Implementation

To carry out the research, four CNN models were employed for evaluating the performance of the models for classification of the imges. The images were classified into 3 categories such as reversed, normal, and corrected. The CNN model selected for the classification study is thus AlexNet, VGG16, ResNet50, and Inception. Each model has unique characteristics that contribute to an in-depth comparison of performance in large and complex dataset.

2.2.1. AlexNet

AlexNet has five convolutional layers, three fully connected layers, and built-in max-pooling operation in a hierarchical structure. It makes use of the ReLU activation function for non-linear activation and accelerating learning over the conventional functions such as sigmoid. Overlapping max pooling is a major innovation in AlexNet, minimizing spatial hassles while preserving important features to avoid overfitting. Also, dropout is used in the fully connected layers for regularization, adding to generalization by avoiding neurons randomly during training[9].

For example, the Adam algorithm was used to train model optimization, and its learning rate was an initial value of 0.001 with a batch size of 32. Implemented the Cross-Entropy Loss function for classification task. We use a pre-specified stopping criterion to quit the training whenever loss is not improving in evaluation phase, which will prevent over-fitting opportunities. Image classification performance of AlexNet was assessed using accuracy, precision, recall, F1-score and area under the curve (AUC) metrics. These metrics also pointed out certain other important aspects in the detection of patterns by the model in learning disabilities.

2.2.2. VGG16

VGG-16 (Visual Geometry Group, University of Oxford) was proposed by Simonyan and Zisserman in 2014 as one of the deep network architectures known for its pure simplicity (convolutional filters mostly with 3 x 3 filter size preferred over other filter sizes). Considering this property, it has an amazingly high capacity to learn. The model uses a deep CNN with many layers that extract features and patterns from input data at different levels.VGG-16 is a well-known model that works well for detecting

learning disability due to the ability of understanding and capturing complex patterns from a large dataset. It can understand features across various handwriting samples.

VGG-16 is a complex image classification model that has 13 convolutional layers and 3 full connected layers. It uses ReLU activation to speed up learning and 3x3 filters in all convolutional layers to maintain a uniform design [10].

The Adam optimization algorithm was used, with an initial learning rate of 0.001 to allow for efficient weight updates. A batch size of 32 was used to maximize training performance with minimal memory consumption. The cross-entropy loss function was used in classification. An early stopping point was set to avoid overfitting, which halted training when additional validation loss reduction was not seen.

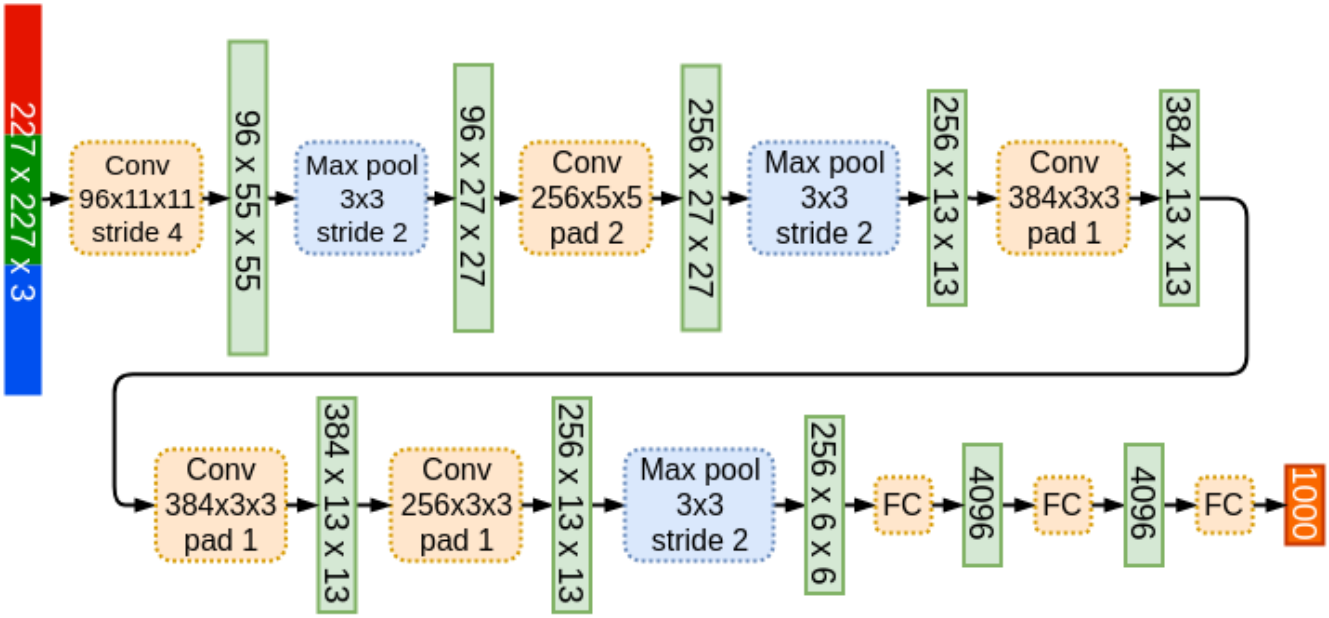

Fig 2. Architecture of AlexNet

Fig 3. Architecture of VGG 16

2.2.3. ResNet-50

ResNet-50, a shortened form of Residual Network with 50 layers, was established by He et al. in 2015 and has greatly revolutionized the sector of deep learning by efficaciously addressing the problem of vanishing gradients in ultra-deep neural models. Its groundbreaking use of residual connections allows ResNet-50 to learn deeper architectures with increased efficiency, making it a powerful tool for complicated tasks of classification. With its ability to differentiate fine features and patterns, ResNet-50 finds it easy to predict learning disabilities based on the examination of fine traits within varied datasets [11]. The structure of ResNet-50 has 50 layers that encompass convolutional, batch normalization, and fully connected layers, with a unique feature of residual blocks.

Pre-trained parameters from ImageNet were fine-tuned with the learning disability dataset, thus taking advantage of the pre-trained feature extraction properties of the network for domain-specific task implementations. The images were reformatted to sizes 224×224×3 pixels to conform to ResNet-50 input dimension requirements.

Used Adam optimizer and started with a start off learning rate 0.0001, to ascertain efficiency and consistent updates of model weights. Adopted a 32 batch size in order to achieve the right balance between learning stability and computational cost. Used the cross-entropy loss function for the task of classification. Applied the step decay learning-rate scheduler to decrease the learning rate in a systematic manner as training continued, thus promoting convergence. An early stopping mechanism was implemented early on to prevent training once the validation loss had plateaued, in order to prevent overfitting.

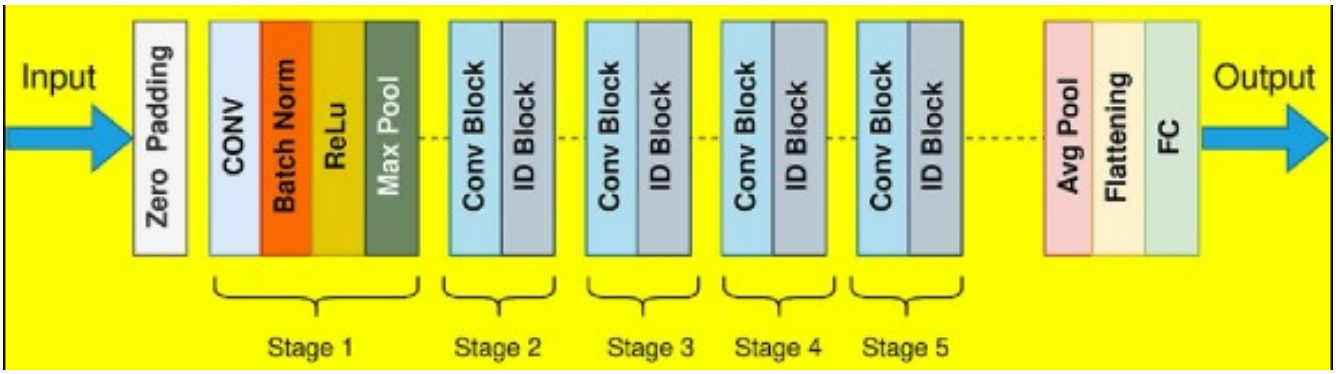

Fig 4. Architecture of ResNet-50

2.2.4. Inception

The Inception architecture, commonly reffered to as GoogLeNet, was introduced by Szegedy et al. in 2015 as a key contribution of the Google team to Deep Learning. The architecture is a leading development in CNNs by the application of "Inception modules," which allow for effective extraction of spatial and channel-wise features at multiple scales. Due to its ability to process intricate data while maintaining computational efficiency, Inception architecture is especially beneficial for uses like predicting learning disabilities, where data usually includes complex patterns and subtle variations.

To modify the Inception architecture for learning disability prediction, a number of changes were made: The final layer, originally fully connected, was replaced for the particular task that was set to make predictions in line with the different learning disability categories. The pre-trained paramenters from ImageNet were fine-tuned in the learning disability dataset, thus leveraging the architecture's superior feature extraction for domain-specific tasks. Images were rescaled into sizes of 299×299×3 pixels to match the Inception architecture's input dimension specifications [12].

2.3 Overview of stacked ensemble model

The proposed methodology employs a stacked deep ensemble based learning architecture to enhance the accuracy and robustness of learning disability detection based on handwriting patterns. Ensemble learning strategies merge the best features of different models to improve generalization and performance, and stacking—one of the most effective ensemble techniques—facilitates hierarchical learning by leveraging the outputs or feature representations of base models as inputs to a meta-learner.

In the present study, four pre-trained CNN frameworks—ResNet50, InceptionV3, VGG16, and AlexNet—were initially evaluated for their performance in Attribute extraction. Among these, ResNet50 and InceptionV3 demonstrated outstanding classification accuracy performance and generalization and were therefore selected as the base models for the ensemble.

These selected models were used to select deep features from the input images, which represent diverse perspectives due to their differing architectural characteristics. ResNet50 incorporates deep residual connections that preserve critical low-level and high-level features across layers, while InceptionV3 utilizes multi-scale convolutional operations that are well-suited to capturing spatial hierarchies. The feature maps generated by each model were subjected to global average pooling and then concatenated to form a unified feature representation.

Next, the merged feature vector was introduced to a fully connected neural network serving as the meta-learner. The meta-learner was trained to gain the optimal mapping from the combined deep features to the final output class, representing one of the handwriting categories linked to learning disabilities (e.g., correct, casual, or reverse). The selection of stacked architecture

allows the model to leverage both the depth of ResNet and the scale-awareness of Inception, resulting in a more holistic and accurate prediction system.

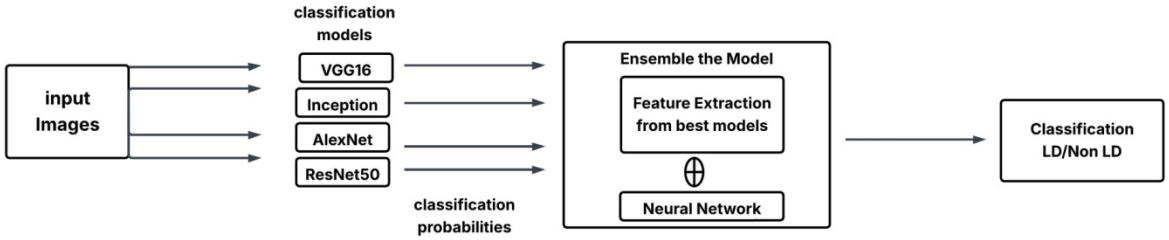

Fig 5. Flow Diagram of Stacked Ensemble Model

The proposed stacked ensemble approach overcomes the challenges and limitations of single individual models. The proposed method makes use of combined streghth of multiple models to enhance the classification performance. We have used 50000 handwritten images for the study that differ in their style and contain minute details also. So, the approach is able to detect the learning disability in all the conditions that can vary greatly and frequently contains subtle details. Hence a method is needed that can identify learning disability in any circumstances. This approach works well in identifying learning disabilities from different types of handwriting with various patterns and structures. It helps the model understand both the fine details and the overall form of the writing.

Algorithm 1: Handwriting Classification using Stacked CNN Ensemble
Input: Dataset D of handwritten images
Output: Predicted class label for new images
1: Preprocess each image in D
  1.1 Resize image to 224 × 224 pixels
  1.2 Normalize pixel values
  1.3 Apply model-specific preprocessing
2: Load pre-trained CNN models
  2.1 ResNet50 ← LoadModel(exclude top layer)
  2.2 InceptionV3 ← LoadModel(exclude top layer)
3: For each image ∈ D do
  3.1 resnet_features ← GlobalAveragePooling2D(ResNet50(image))
  3.2 inception_features ← GlobalAveragePooling2D(InceptionV3(image))
  3.3 combined_features ← Concatenate(resnet_features, inception_features)
End For
4: Define meta-learner neural network
  4.1 Dense layer with ReLU activation
  4.2 Dropout layer to prevent overfitting
  4.3 Output layer with Softmax activation for classification
5: Train meta-learner using combined_features
6: Evaluate model performance using:
  6.1 Accuracy
  6.2 Precision
  6.3 Recall
  6.4 F1-Score
7: Return predicted class label for new handwritten images

3 Results and Discussion

3.1 Evaluation Metrics

For performance assessment of the Convolutional Neural Network (CNN) models, specifically AlexNet, VGG16, Inception, and ResNet50, a range of standard evaluation metrics were employed. The same evaluation matrix is used to assess the implemented stacked ensemble model too. The definitions of these evaluation metrics utilized for the assessment of the CNN models are mentioned as follows:

To evaluate the performance of the CNN models—AlexNet, VGG16, Inception, and ResNet50— a range of standard evaluation metrics were used. The main metrics applied for assessing these models are explained below:

- **Accuracy**

The formula for accuracy is given by

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

True Positives (TP): Cases where the model correctly predicts a positive result.

True Negatives (TN): Cases where the model correctly predicts a negative result.

False Positives (FP): Cases where the model wrongly predicts positive when it is actually negative.

False Negatives (FN): Cases where the model wrongly predicts negative when it is actually positive.

- **Precision**

The formula for precision is given by

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

- **Recall**

The formula for recall is given by

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

- **F1-Score**

The formula for the F1-Score is given by

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

- **AUC-ROC**

The formula for Area Underthe Receiver Operating Characteristic Curve (AUC-ROC) is given by

$$AUC-ROC = \int_0^1 TPR(FPR)d(FPR) \tag{5}$$

Where:

TPR stands for the True Positive Rate, and FPR corresponds to the False Positive Rate.

Table 1. Performance Metrics for CNN transfer learnig Models

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
AlexNet	0.83	0.82	0.83	0.825	0.86
VGG 16	0.88	0.86	0.87	0.865	0.89
Inception	0.91	0.89	0.9	0.895	0.91
ResNet-50	0.94	0.93	0.94	0.935	0.96

3.2 Results of CNN Transfer Learning Models

Among the evaluated models, ResNet-50 performed the best. It achieved around 94% accuracy and showed strong precision and recall, along with an AUC-ROC of 96%. Its residual connections help preserve important information across deep layers, making it particularly good at capturing complex handwriting patterns.

Inception followed closely, reaching about 91% accuracy. Its parallel convolutional layers allow it to detect features at multiple scales, which help in generalizing across different handwriting styles.

VGG-16, though simpler, still performed well with roughly 88% accuracy and captured detailed features using its deep layers with small receptive fields. AlexNet, the oldest architecture tested, showed moderate performance around 85–89% but still demonstrated potential for feature learning relevant to learning disability classification.

Overall, ResNet-50's deep residual structure clearly gives it an edge, while the other models offer varying strengths in multi-scale or detailed feature extraction.

The comparison of the four convolutional neural network (CNN) models shows uniform trends in the different evaluations. ResNet-50 consistently shows better performance in all the metrics, making it the best model for predicting learning disabilities. Inception is still the second most efficient model, showing good performance, though not as good as ResNet-50. VGG-16 shows a balanced performance, which is better than AlexNet but not as strong as the robustness shown by Inception and ResNet-50. AlexNet recorded the least favorable performance in the comparison, positioning it at the lower end of the model hierarchy. On the other hand, ResNet-50 emerged as the top-performing architecture, with the results reflecting a balanced and realistic assessment of its superiority

1. **Precision and Recall Evaluation**

   Precision measures the accuracy of Identified positive instances and recall measures the Proportion of true positive instances correctly identified. ResNet-50 model had the highest precision value (0.93) and recall value (0.94), reflecting its ability to minimize false positives while identifying a large proportion of true positive cases. With precision of 0.89 and recall of 0.90, Inception model reflected very good predictive power. VGG-16 produced marginally lower but still competitive performances, recording precision of 0.86 and recall of 0.87, indicating that it makes accurate predictions, albeit marginally less so. On the other hand, AlexNet showed the weakest results, with the lowest precision (0.82) and recall (0.83), which reflected its overall poorer performance in classification.

2. **F1-Score and AUC-ROC Analysis**

   The F1-score is an important metric because it balances precision and recall, giving an overall measure of how well a model performs. Among the models, ResNet50 had the best F1-score of 0.935, showing the strongest balance. Inception came next with 0.895, followed by VGG16 at 0.865. AlexNet scored the lowest at 0.825, reflecting its weaker performance compared to the others.

By analyzing the AUC-ROC metric it is evident that all the used models perform well in classifying the images into different classes. ResNet50 stands out by scoring 0.96 showing outstanding performance in classification.Inception shows strong ability of the model by hitting 0.91 score while VGG16 and AlexNet had lowest score of 0.89 and 0.86 respectively. The results shows that ResNet50 and Inception models are appropriate for defining class boundaries, but the other two models are less effective for classification.

3.3 Stacked Ensemble Model Performance

The ensemble stacking method was applied by combining Inception and ResNet-50, identified as the highest-performing CNN models based on their individual performance scores. ResNet-50 achieved the top accuracy of 94%, followed by Inception with 91% accuracy. Stacking leverages the strengths of various models to build more powerful and generalized prediction framework.

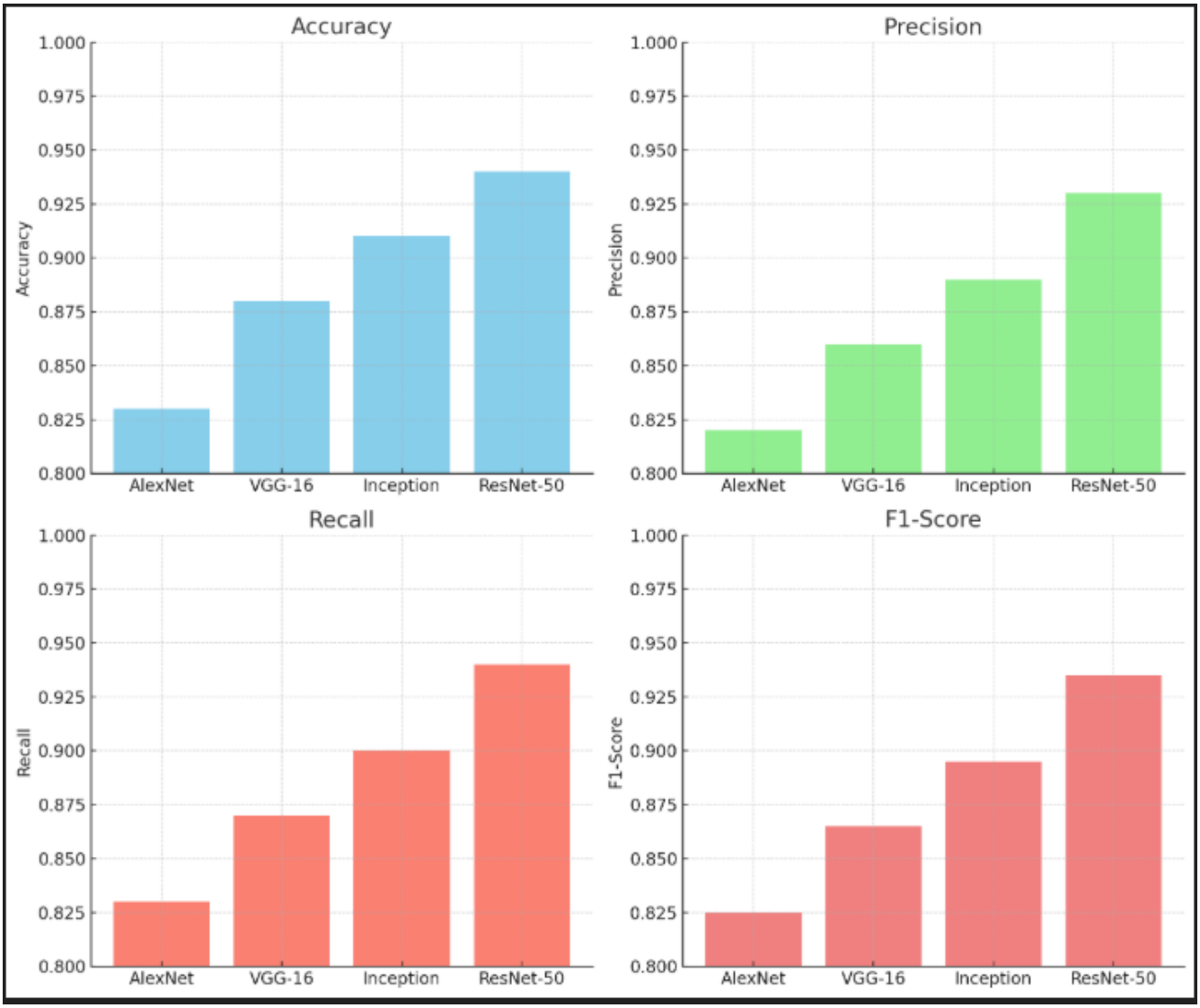

Fig 6. Performance of the CNN models (AlexNet, VGG-16, Inception, and ResNet-50)

The ensemble framework utilized feature extractors from Inception and ResNet-50, where the derived features were concatenated and input into a meta-learner implemented as a fully connected neural network. This meta-model was trained to learn the most effective combination of outputs from the base models, facilitating bias reduction and variance minimization to enhance overall prediction accuracy.

In the learning epoch, the model was tested for over 30 epochs, and its performance was monitored carefully using prominent indexes: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. During training and testing, accuracy had natural fluctuations—characteristic of learning and generalization—but had a sure upward trend, reaching a peak accuracy of about 95%. This minor fluctuation is to be anticipated in deep learning models, particularly when learning intricate patterns from feature spaces. Also, both the training and testing losses showed a declining trend with occasional plateaus before stabilizing at lower levels, indicating that the model was successfully reducing the error without overfitting.

The final performance of the stacked ensemble model surpassed all individual base models, affirming the advantage of stacking in enhancing model robustness and predictive power. This validates the efficacy of ensemble strategies in tasks involving complex data, such as learning disability detection from handwriting patterns.

/*Input: Handwritten image dataset D
Output: Predicted learning disability class (Correct, Casual, Reverse)*/

1. Preprocess the input images:
   - Resize images to 224x224
   - Normalize pixel values
   - Apply model-specific preprocessing

2. Load pre-trained CNN models:
   - Load ResNet50 (exclude top layer)
   - Load InceptionV3 (exclude top layer)

3. For each image in D:
   a. Extract deep features using ResNet50:
      resnet_features ← GlobalAveragePooling2D(ResNet50(image))

   b. Extract deep features using InceptionV3:
      inception_features ← GlobalAveragePooling2D(InceptionV3(image))

   c. Concatenate features:
      combined_features ← Concatenate(resnet_features, inception_features)

4. Feed combined features into neural network (meta-learner):
   a. Dense Layer (ReLU activation)
   b. Dropout (to prevent overfitting)
   c. Output Layer (Softmax for classification)

5. Train the meta-learner using training data
6. Evaluate the model on test data using metrics: accuracy, precision, recall, F1-score

Return: Predicted class label for new handwritten inputs

Fig 7. Algorithm of stacked ensemble model

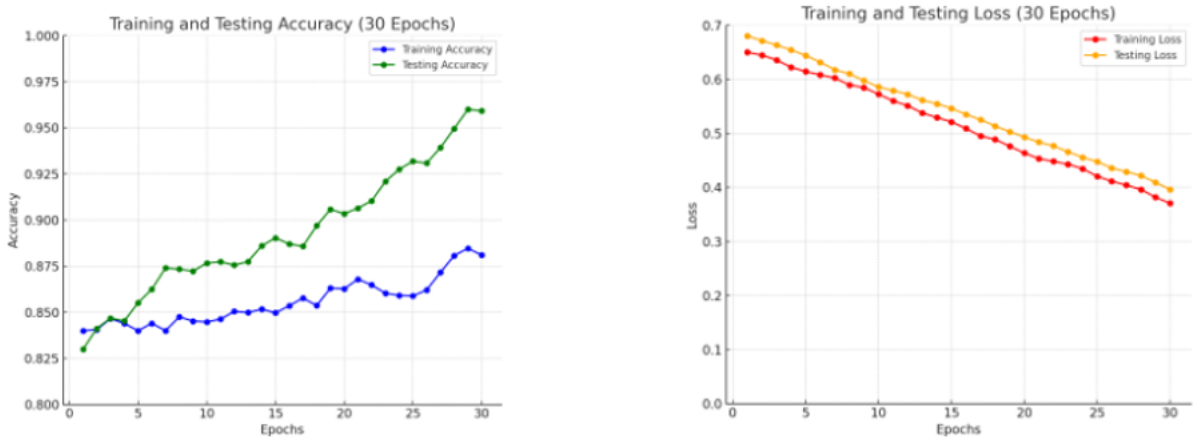

Fig 8. Accuracy and Loss of proposed model (Stacked ensemble model)

3.4 Performance Comparison with Existing Models

The table provides a comparative evaluation of different classification models and technique, highlighting their respective accuracies. The proposed method evaluates multiple deep learning models, with ResNet50 emerging as the best performer, shows an accuracy of 95%. This indicates the functionality of residual connections in addressing the vanishing gradient problem and improving feature extraction. In contrast, Inception and VGG16 achieved slightly lower accuracies of 91% and 88%, respectively, while AlexNet lagged behind at 85%. The superior performance of ResNet50 suggests that deeper architectures with optimized feature extraction techniques yield better classification results.

Table 2. Comparison with existing models

Reference	Classification Method / Model	Accuracy
Proposed Method	AlexNet	0.85
	VGG16	0.88
	Inception	**0.91**
	ResNet50	**0.94**
	Stacked ensemble model	**0.95**
Royan et al.[13]	SVM	0.76
	CNN	0.89
Hu et al.[14]	Multi-scale transformer	0.89
Sharma et al.[15]	CNN sequential method	0.94
S. Saed et al.[16]	MCNN-14	0.93

The proposed study demonstrates that ResNet-50 outperforms traditional and earlier deep learning approaches in learning disability classification, achieving 94% accuracy, with the stacked ensemble model further improving it to 95%. Royan et al.[13] reported a maximum accuracy of 88.9% with CNN, while SVM reached 76%. Their study also highlighted the limitations of traditional feature-based methods. They state that SVM is better for faster training process and better model for resource limited prediction process.

Similarly, Hu et al.[14] used Multi-Scale Transformer model achieved 89% accuracy by incorporating multi-scale feature fusion mechanism and optimizing the self-attention mechanism, comparable to the ResNet50, ResNext, ViT and VGG19 models in this study. Sharma et al.[15] observed a performance of 94 % with sequential CNN method.They used content based image retrieval techniques to retrieve images from database and then added the CNN algorithm to detect different features and classified the images into 3 different classes. S. Saed et al.[16] demonstrated multiple-CNN architecture attained a classification accuracy of 93.08%, offering autonomous learning and enhanced feature extraction capabilities to enhance model performance.

Compared to these recent studies, the proposed stacked ensemble leverages complementary strengths of multiple CNN architectures, outperforming both single-model and hybrid CNN approaches. Models like VGG16 and Inception still perform well but combining them with ResNet-50 in an adaptive ensemble clearly improves accuracy and generalization. These results suggest that ensemble-based deep learning can be more effective for early detection of learning disabilities. It also appears that further gains could be made by blending hybrid strategies with deep residual architectures.

4 Conclusion

This study indicates ResNet-50, and inception models achieved the best results across major evaluation metrics. ResNet-50 is the most effective CNN model for predicting learning disabilities with accuracy around 94% and strong precision and recall. Inception was close behind, reaching about 91% and showing steady performance across tasks. Combining these models in a stacked ensemble with a meta-level neural network pushed accuracy to 95%, while also balancing precision and recall. This suggests that merging complementary models helps capture subtle handwriting features and provides a practical approach for early detection of learning disabilities in schools.

The ensemble outperformed both base model in terms of accuracy, robustness, and generalization by combining the advantages of both CNN models-ResNet-50 and Inception and then for harnessing their outputs using a meta-level neural network for better performance. The potential of this approach is that they achieved a balanced trade-off between precision and recall, making it a practical solution for real-world educational screening.

Future research could concentrate on enhancing the stacked ensemble approach by incorporating contemporary architectures like DenseNet and EfficientNet and selecting adaptive learning techniques to improve training outcomes. Using

explainable AI (XAI) techniques is another helpful strategy that would improve the models' transparency and offer more understandable insights into their predictions. The research could be taken into another level by implementing a system that can be used in real time. Most of the school teachers or parents may have minimal configuration mobile phones or laptops. It is very difficult to predict the learning disability by using high end laptops or desktops. So future research can concentrate on developing a lightweight system to predict learning disability by using normal systems with minimal configurations.

5 Acknowledgement

The authors thank, DST-FIST, Government of India for funding towards infrastructure facilities at St.Joseph's College (Autonomous), Tiruchirappalli - 620 002.

References

1) Prabhala JC, Ragoju R, Kupilli V, Chesneau C. Enhanced Early Detection of Dysarthric Speech Disabilities Using Stacking Ensemble Deep Learning Model. *Machine Learning with Applications*. 2025;p. 100721. https://doi.org/10.1016/j.mlwa.2025.100721.
2) Soumya K, Charles PJ, Kavitha S. Comparative Analysis of Convolutional Neural Network Transfer Learning Models for Predicting Learning Disability. *Journal of Electrical Systems*. 2024;20(3). Available from: https://journal.esrgroups.org/jes/article/view/6455/4506.
3) Mayfield JD, et al. Time-Dependent Deep Learning Prediction of Multiple Sclerosis Disability. *Journal of Imaging Informatics in Medicine*. 2024;p. 1–19. 10.1007/s10278-024-01031-y.
4) Vajs I, et al. Accessible dyslexia detection with real-time reading feedback through robust interpretable eye-tracking features. *Brain Sciences*. 2023;13(3):405. https://doi.org/10.3390/brainsci13030405.
5) Kulasekara KMG, Silva DGDHD, Pranavan S. Enhancing Sinhala Writing Skills in Dyslexic Students through Personalized and Gamified Learning. 2025. Doctoral dissertation. Available from: https://dl.ucsc.cmb.ac.lk/jspui/handle/123456789/4901.
6) of Standards NI, Technology. NIST Special Database 19: Handprinted Forms and Characters Database. Available from: https://www.nist.gov/srd/nist-special-database-19.
7) Patel S. A-z handwritten alphabets in csv format. Available from: https://www.kaggle.com/sachinpatel21/az-handwritten-alphabets-in-csv-format.
8) Salaheddine FI. Convolutional neural networks (CNN): Models and algorithms. IGI Global. 2023. Available from: https://www.igiglobal.com/viewtitlesample.aspx?id=316788&ptid=296570&t=convolutional+neural+networks+(cnn)%3a+models+and+algorithms.
9) Verdhan V. VGGNet and AlexNet networks. Berkeley, CA. Apress. 2021. https://doi.org/10.1007/978-1-4842-6616-8_4.
10) Raghuvanshi S, Dhariwal S. The VGG16 Method Is a Powerful Tool for Detecting Brain Tumors Using Deep Learning Techniques. *Engineering Proceedings*. 2023;59(1):46. https://doi.org/10.3390/engproc2023059046.
11) Ali RR, Yaacob NM, Alqaryouti MH, Sadeq AE, Doheir M, Iqtait M, et al. Learning architecture for brain tumor classification based on deep convolutional neural network: Classic and ResNet50. *Diagnostics*. 2025;15(5):624. https://doi.org/10.1007/978-1-4842-6168-2_6.
12) Gursel AT, Kaya Y. Mam-Incept-Net: a novel inception model for precise interpretation of mammography images. *PeerJ Computer Science*. 2025;11:e3149. https://doi.org/10.7717/peerj-cs.3149.
13) Royan YI, Pramono P, Asri AAK. Performance Comparison of Convolutional Neural Networks (CNN) and Support Vector Machine (SVM) Algorithms in Human Face Classification. *G-Tech: Jurnal Teknologi Terapan*. 2025;9(3):1544–1553. https://doi.org/10.70609/g-tech.v9i3.7384.
14) Hu J, Xiang Y, Lin Y, Du J, Zhang H, Liu H. Multi-scale Transformer architecture for accurate medical image classification. 2025. https://doi.org/10.1145/3730436.3730505.
15) Sharma A, Phonsa G. Image Classification Using CNN. 2021 April. Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021. Available from: https://ssrn.com/abstract=3833453. http://dx.doi.org/10.2139/ssrn.3833453.
16) S S, B T, K K, A SM. An Efficient Multiple Convolutional Neural Network Model (MCNN-14) for Fashion Image Classification. Tehran, Iran, Islamic Republic of. 2024. https://doi.org/10.1109/ICWR61162.2024.10533341.