The results of this study highlight the comparative performance of various deep learning models, including CNN, CNN with an attention layer, ResNet50, MobileNetV2, and DenseNet121, in classifying brain tumors into four distinct classes: glioma, meningioma, nontumor, and pituitary. Each model was evaluated using precision, recall, and F1-score to assess their effectiveness in identifying and differentiating between these tumor types. The baseline CNN model demonstrated strong performance overall, particularly in classifying nontumor and pituitary tumors, with F1 scores reaching 0.98. However, it exhibited slightly reduced accuracy in identifying glioma and meningioma tumors. Introducing an attention mechanism to the CNN architecture improved performance, particularly in the meningioma class. This indicates that the attention layer helped the model focus on more relevant features, thus enhancing classification accuracy. This section provides a detailed analysis of each model's performance, offering insights into their strengths and limitations in handling this critical medical imaging task.

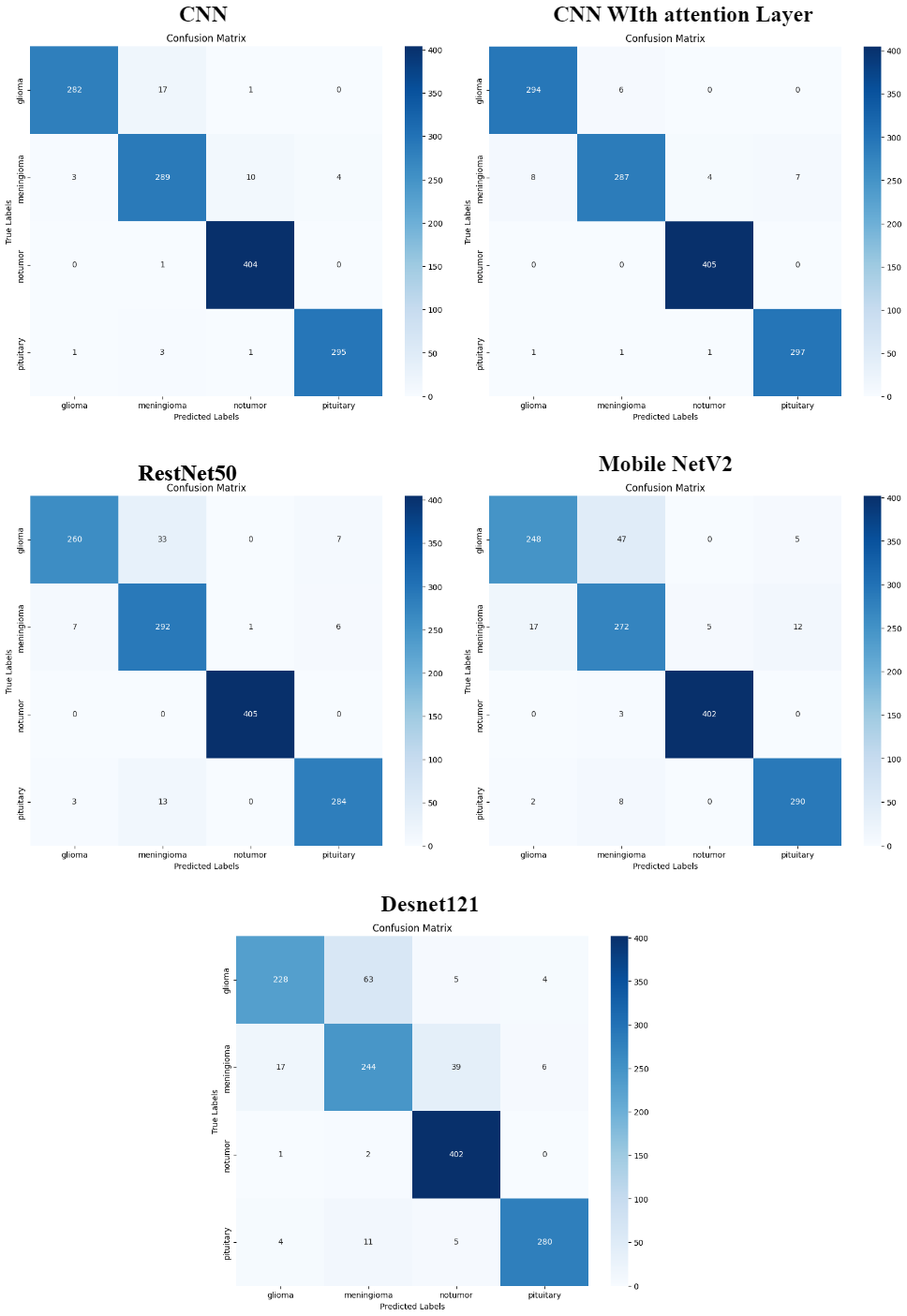
First of all, the class-wise results are analyzed. Each class's precision, recall, and f1 score on each applied model are evaluated individually. The baseline CNN performed well in nontumor and pituitary classifications (F1-scores of 0.98) but was slightly less accurate for glioma and meningioma. Adding an attention layer improved CNN's performance, particularly for meningioma, increasing the F1-score to 0.96. ResNet50 excelled in nontumor classification but struggled with glioma and meningioma, while MobileNetV2. Despite being lightweight, it showed lower accuracy for these classes. DenseNet121 had the lowest overall performance, particularly for glioma and meningioma, indicating a need for further model refinement or larger datasets. The consistently high performance in nontumor classification suggests this class is more easily distinguished, while glioma and meningioma remain challenging. The attention-enhanced CNN and ResNet50 were the most balanced models, but further improvements, such as ensemble learning or data augmentation, may enhance performance in complex cases. This analysis highlights the need for careful model selection and tuning based on task-specific challenges.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Target Class | Precision | Recall | F1-score | Support |
| CNN | glioma | 0.99 | 0.94 | 0.96 | 300 |
| meningioma | 0.93 | 0.84 | 0.94 | 306 |
| nontumor | 0.97 | 1.00 | 0.98 | 405 |
| pituitary | 0.99 | 0.98 | 0.98 | 300 |
| Inception CNN | glioma | 0.97 | 0.98 | 0.98 | 300 |
| meningioma | 0.98 | 0.94 | 0.96 | 306 |
| nontumor | 0.99 | 1.00 | 0.99 | 405 |
| pituitary | 0.98 | 0.99 | 0.98 | 300 |
| ResNet50 | glioma | 0.96 | 0.87 | 0.91 | 300 |
| meningioma | 0.86 | 0.95 | 0.91 | 306 |
| nontumor | 1.00 | 1.00 | 1.00 | 405 |
| pituitary | 0.96 | 0.95 | 0.95 | 300 |
| MobileNetV2 | glioma | 0.93 | 0.83 | 0.87 | 300 |
| meningioma | 0.82 | 0.88 | 0.86 | 306 |
| nontumor | 0.99 | 0.99 | 0.99 | 405 |
| pituitary | 0.94 | 0.97 | 0.96 | 300 |
| DenseNet121 | glioma | 0.91 | 0.76 | 0.83 | 300 |
| meningioma | 0.76 | 0.80 | 0.78 | 306 |
| nontumor | 0.89 | 0.99 | 0.94 | 405 |
| pituitary | 0.97 | 0.93 | 0.95 | 300 |

The overall results of these models show a clear hierarchy in performance, with CNN with the attention layer leading the group, achieving the highest Precision, Recall, and F1-score of 0.98, indicating its superior ability to identify and classify brain tumor images correctly. The baseline CNN follows closely with uniformly strong performance across all metrics and has 0.97 scores, making it a robust choice, though slightly less effective than Inception CNN. ResNet50, while still competent, shows a slight drop in performance. The RestNet has a 0.94 score on Recall and F1 Score, which may indicate some challenges in capturing all relevant cases accurately. MobileNet20, known for its lightweight architecture, exhibits slightly lower scores, with 0.92 across all metrics, suggesting a trade-off between efficiency and accuracy. DenseNet12 shows the lowest performance, with an F1-score of 0.87, reflecting potential difficulties in complex feature extraction and indicating it may not be as reliable for this specific classification task. Overall, while all models perform reasonably well, Inception CNN is the most effective, with DenseNet12 requiring further refinement or possibly using additional data to improve its classification capabilities.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Precision | Recall | F1-score |
| CNN | 0.97 | 0.97 | 0.97 |
| Inception CNN | 0.98 | 0.98 | 0.98 |
| ResNet50 | 0.95 | 0.94 | 0.94 |
| MobileNet20 | 0.92 | 0.92 | 0.92 |
| DenseNet12 | 0.88 | 0.87 | 0.87 |

The performance of all applied deep learning models was further evaluated by plotting the confusion matrices for each model. These confusion matrices, visualized in Figure 1, provide a detailed view of each model's classification accuracy and errors. Additionally, the training history graphs for all models were plotted to assess their generalization capabilities. This comprehensive evaluation through confusion matrices and history graphs offers valuable insights into the strengths and weaknesses of each model.



The history graph of each model, as shown in Figure 2, reveals that the CNN model with an attention layer exhibits a more consistent and smooth increase in accuracy compared to the other models applied in this study. This smoothness in the accuracy curve indicates better generalization, suggesting that the CNN with attention is less prone to overfitting and maintains stable performance across different data distributions.

