Automated Classification of Prostate Cancer Severity Using Pre-trained Models

Sílvia Barros[1], Vitor Filipe[1,2], and Lio Gonçalves[1,2(✉)]

[1] School of Science and Technology, University of Trás-os-Montes e Alto Douro (UTAD), 5000-811 Vila Real, Portugal
silviab477@outlook.pt, {vfilipe,lgoncalv}@utad.pt
[2] INESC Technology and Science (INESC TEC), 4200-465 Porto, Portugal

Abstract. Prostate cancer is one of the most common types of cancer in men. The ISUP grade and Gleason Score are terms related to the classification of this cancer based on the histological characteristics of the tissues examined in a biopsy. This paper explains an approach that utilizes and evaluates pre-trained models such as ResNet-50, VGG19, and InceptionV3, regarding their ability to automatically classify prostate cancer and its severity based on images and masks annotated with ISUP grades and Gleason Scores. At the end of the training, the performance of each trained model is presented, as well as the comparison between the original and predicted data. This comparison aims to understand if this approach can indeed be used for a more automated classification of prostate cancer.

Keywords: ISUP Grading · Gleason Score · Resnet-50 · VGG19 · InceptionV3

1 Introduction

Prostate cancer is currently one of the main cancers fought by the male population, being the most common in terms of occurrence and the second leading cause of death. In Portugal, approximately 6000 cases of prostate cancer are diagnosed each year, which accounts for twenty one percent of all cancers in males. This translate to an average of one in six individuals who will be diagnosed with this cancer in their lifetime, This leads us to conclude that in Portugal, as well as in other countries, prostate cancer is considered a significant issue. In order to study this cancer, a methodology was developed by Donald Gleason in 1960 [1].

The Gleason grading system is one of the most important factors in clinical decision-making for prostate cancer patients [2]. It is entirely based on the classification of tumor growth patterns, and in recent years, it has become clear that some individual growth patterns have independent prognostic value and can be used to improve personalized risk stratification. By providing a standardized and objective classification of tumor growth patterns, this grading system enables

A. Cunha et al. (Eds.): MobiHealth 2023, LNICST 578, pp. 469–482, 2024.
https://doi.org/10.1007/978-3-031-60665-6_35

physicians to assess disease progression and determine the optimal therapeutic approach for each patient, making it an important tool in the field of urology.

The Gleason score is a tissue-based prognostic marker for prostate cancer that suffers from variability, leading to inconsistencies and classification errors. As a result, artificial intelligence (AI) systems based on deep learning have emerged, showing promising results in achieving performance similar to that of pathologists in Gleason grading [4].

Over the years, the Gleason grading system has been enhanced with the use of artificial intelligence (AI). AI has been applied to the analysis of histopathological images from prostate biopsy samples, assisting pathologists in the classification of tumor growth patterns. In this regard, AI algorithms can be trained on large datasets containing histological images previously analyzed by experts. These algorithms can learn to identify and classify the different patterns of prostate cancer growth, providing automated and accurate evaluation. The integration of artificial intelligence into this grading system brings significant benefits, which can reduce inter-observer variability in the interpretation of growth patterns and improve result consistency.

The ISUP (International Society of Urological Pathology) classification is a grading system used to assess prostate cancer based on histological features observed in tissue samples examined by pathologists. This classification was developed by the International Society of Urological Pathology and is widely used to provide prognostic information and assist in treatment planning.

AI systems are trained on large datasets that contain information about tissue characteristics and patterns associated with different disease grades, and they can learn to automatically recognize and extract these patterns. This assists experts in accurately classifying the disease grade. Additionally, AI systems can be used to create prediction models that estimate the risk of disease progression based on Gleason and ISUP grading.

Wouter Bulten et al. [3] evaluated the impact of AI assistance on the diagnostic performance of pathologists in Gleason grading of prostate biopsies. The results showed that AI assistance significantly improved the performance of pathologists, reducing variability and leading to more consistent grading. The AI system outperformed most pathologists in the unassisted read, but the panel's performance surpassed that of the AI system in the assisted read. The study highlights the potential of AI systems to support pathologists in achieving higher accuracy and consistency in Gleason grading, particularly in regions with limited access to specialized pathologists. Further research and validation are needed to fully understand the benefits of AI assistance in clinical practice.

Wouter Bulten et al. [4] created the PANDA challenge. The PANDA challenge aimed to improve the Gleason grading of prostate cancer by using AI algorithms. The challenge organized a global competition and provided a large dataset for algorithm development and validation. The dataset included over 12,000 prostate biopsy images from different sites. Experienced uropathologists established reference standards for the training and validation sets, ensuring consistency across different regions. The competition attracted teams from 65

countries, and the algorithms submitted by the teams were blindly validated on an internal validation set. The top-performing algorithms utilized deep learning methods and demonstrated high agreement with the uropathologists. In the external validation sets, the algorithms showed high agreement with the reference standards, although they had a tendency to overdiagnose benign cases as low-grade cancer. The algorithms performed better than general pathologists in terms of sensitivity but had lower specificity in correctly identifying benign cases. Overall, the AI algorithms achieved pathologist-level concordance and outperformed previous methods. However, the study had limitations, such as a limited number of participating teams and a focus on specific cancer types. Further research is needed for broader evaluation and to address potential errors.

According to Kimmo Kartasalo et al. [5], the use of artificial intelligence (AI) systems for diagnosing and grading prostate cancer in biopsies shows promise in overcoming the variability between pathologists. AI systems have demonstrated high accuracy in cancer detection and Gleason grading, comparable to expert pathologists. However, further validation using diverse and independent datasets is needed before clinical implementation. Challenges include the need for larger and more diverse training datasets, the ability to recognize benign mimics of cancer, and the development of anomaly detection systems for quality control. Future developments may involve training AI algorithms using long-term follow-up data for improved prognostication.

According to Lars Egevad et al. [6] Standardization in pathology grading, particularly for prostate cancer, is crucial for accurate diagnosis and treatment decisions. While the Gleason grading system has been widely used, variability among pathologists and evolving definitions have led to disagreements. Pathology Imagebase and AI technology offer opportunities to overcome these challenges by providing a standardized reference and assisting in grading. Collaborative efforts between experts, AI developers, and pathologists are essential to refine and validate these tools, leading to improved consistency and precision in prostate cancer grading, ultimately benefiting patient care.

Felicia Marginean et al. [7] demonstrates the potential of artificial intelligence (AI) in addressing the challenges of Gleason grading variability in prostate cancer diagnosis. The AI algorithm trained on annotated biopsy scans showed high accuracy and reproducibility in detecting cancer and assigning Gleason grades. These holds promise for improving the consistency and efficiency of diagnostic processes, reducing the burden on pathologists, and potentially enhancing patient outcomes. The findings highlight the importance of leveraging AI and medical image analysis to standardize and optimize diagnostic practices in prostate cancer and pave the way for future advancements in AI-assisted pathology.

Liron Pantanowitz et al. [8] focuses on the development and validation of an AI algorithm for prostate cancer diagnosis and Gleason grading. The algorithm exhibited high accuracy in distinguishing between low-grade (Gleason score 6 or ASAP) and high-grade (Gleason score 7–10) tumors, as well as detecting the presence of Gleason pattern 5. The algorithm's performance surpassed previous studies and addressed the limitations of narrow AI approaches. Its successful

deployment in routine clinical practice has the potential to improve diagnostic accuracy, reduce interobserver variability in Gleason grading, and provide valuable insights for treatment planning in prostate cancer.

This paper highlights the training of a network that aims to detect and classify the severity of prostate cancer in images of prostate tissue samples that have been previously analyzed and scored by pathologists according to the Gleason system, which is then converted into an ISUP grade.

Therefore, this paper is divided into sections such as Materials and Methods, Results, Discussion and Conclusion. All sections are written with the aim of providing the reader with a clear explanation of the topic.

2 Material and Methods

2.1 Dataset

To carry out the research, a database provided by the Karolinska Institute and the Radboud University Medical Center was used. It includes:

- A folder containing 10,616 images in TIFF format, with 5,455 images provided by the Karolinska Institute and 5,060 images by the Radboud University Medical Center;
- A folder containing segmentation masks that show which parts of the image contributed to the ISUP grade. Not all training images have label masks, and there may be false positives or false negatives in the masks. These masks are provided to assist in the development of strategies for selecting the most useful image subsets, and their values depend on the data provider; Considering this, the number of images without labeled masks was checked, and a new dataframe was created that removes the rows from the training file where masks are missing.
- A CSV file containing information, such as the ISUP grade and the Gleason Score of each image, corresponding to the training data (train.csv). Their distribution is shown in the Figs. 1.
- A folder with pre-trained models, including ResNet-50, VGG19 and InceptionV3.

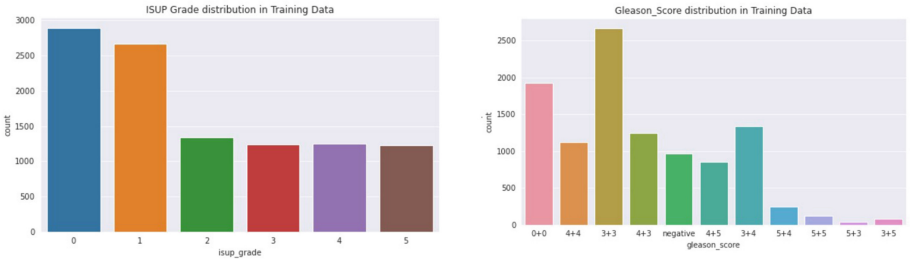

Fig. 1. ISUP Grade and Gleason Score Distribution in Training Data

	isup_grade	image_id
0	0	2873
1	1	2616
2	2	1340
4	4	1245
3	3	1226
5	5	1215

	gleason_score	image_id
1	3+3	2616
0	0+0	1925
2	3+4	1340
4	4+3	1226
5	4+4	1122
10	negative	948
6	4+5	842
8	5+4	248
9	5+5	125
3	3+5	80
7	5+3	43

Fig. 2. Clustering and classification of the data through a color gradient

2.2 Training Images and Masks

As previously described, the masks from the utilized dataset were provided to locate the cancer and aid in understanding each grade of the disease. Thus, the images were displayed along with their corresponding masks. This process is highly useful as it helps visualize the histological patterns present in a tissue sample. The Gleason score and ISUP classification are based on the analysis of these patterns, and the display of images allows for the observation of microscopic details and a better interpretation of tissue characteristics. One training image and corresponding mask is shown in Fig. 3.

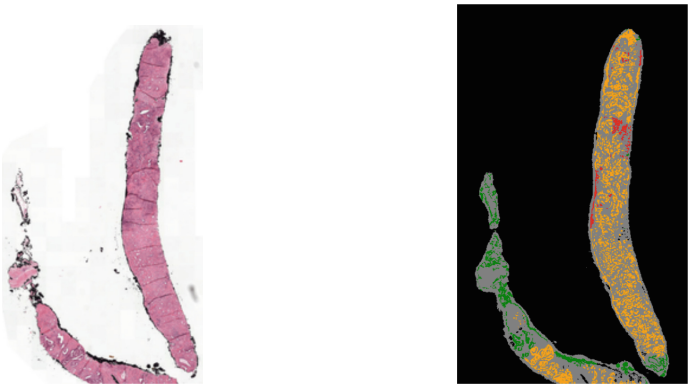

Fig. 3. Image and corresponding mask. Source: Radboud, ISUP:5 and Gleason: 5+4.

2.3 Data Preprocessing

In the Gleason scoring system, there is no ISUP = 2 for a Gleason score of 4+3. Therefore, the existence of images with this classification was verified, and only one was found, which was later removed. After completing this process, the data was clustered to combine the ISUP classification and the Gleason score with the image IDs, as shown in the Fig. 2. The colors present in the tables aim to assist in the analysis and interpretation of the data.

In order to better understand the dataset, the idea arose to replace the "negative" values with "0+0" to have a consistent classification for this value. To achieve this, the data was grouped again, resulting in the graph represented in the Fig. 4. Through analysis, we can observe that the Radboud dataframe does not have "0+0" values, while the Karolinska dataframe does not have "negative" values. Therefore, we conclude that the "negative" values correspond to how Radboud represents "0+0" values. Thus, for the sake of consistency, the "negative" values were replaced with "0+0" (Fig. 5).

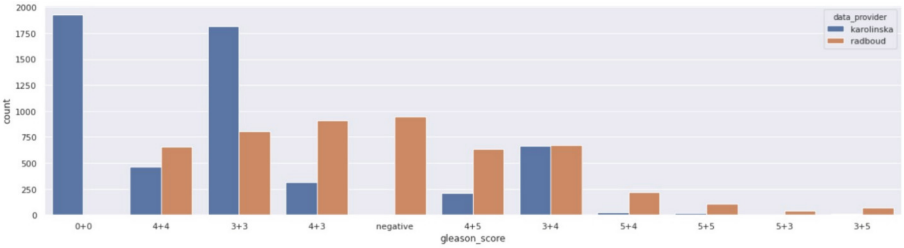

Fig. 4. Count of each Gleason Score value for each data provider institution

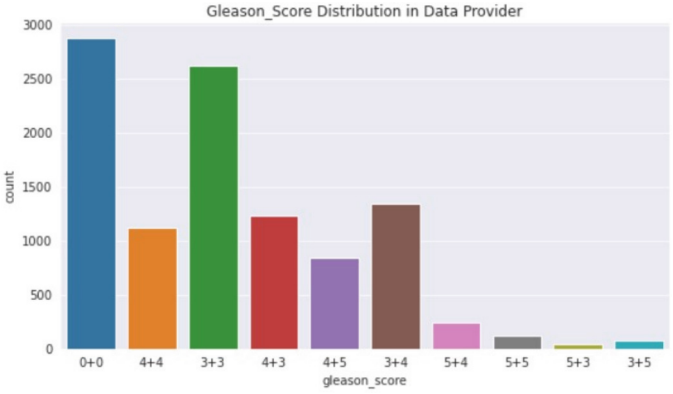

Fig. 5. Gleason Score Distribution after the Replacement

2.4 Gleason Grading and ISUP Classification

The Gleason classification and the ISUP classification are systems used to assess the severity of prostate cancer based on microscopic analysis of tissue samples obtained through biopsy.

The Gleason classification evaluates the appearance of cancer cells in the tissue sample. It consists of a primary Gleason score, which represents the dominant area, and a secondary Gleason score, which represents the second most prevalent area. Both scores range from 1 to 5 and are summed to obtain the total Gleason score, which ranges from 2 to 10. The higher the total Gleason score, the more severe the cancer.

The ISUP classification (International Society of Urological Pathology) is a newer classification system that aims to simplify the interpretation of the Gleason score. It groups Gleason scores into five categories, from 1 to 5. Category 1 corresponds to a Gleason score of 6, category 2 corresponds to Gleason scores of $3+4=7$ or $3+3=6$ with cribriform pattern, category 3 corresponds to Gleason scores of $3+4=7$ or $4+3=7$ without cribriform pattern, category 4 corresponds to Gleason scores of $4+4=8$ or $3+5=8$ or $5+3=8$, and category 5 corresponds to Gleason scores of 9 and 10. The ISUP classification aims to provide a simplified and standardized assessment of the severity of prostate cancer.

The training file in CSV format contains annotations for each image, including both the Gleason and ISUP classifications such as the image ID and the data provider. This can be seen in Fig. 6.

2.5 Resnet-50

The ResNet-50 model is a highly effective approach for image classification tasks, particularly in the field of prostate cancer analysis. By utilizing pre-trained

	image_id	data_provider	isup_grade	gleason_score
0	0005f7aaab2800f6170c399693a96917	karolinska	0	0+0
1	000920ad0b612851f8e01bcc880d9b3d	karolinska	0	0+0
2	0018ae58b01bdadc8e347995b69f99aa	radboud	4	4+4
3	001c62abd11fa4b57bf7a6c603a11bb9	karolinska	4	4+4
4	001d865e65ef5d2579c190a0e0350d8f	karolinska	0	0+0
5	002a4db09dad406c85505a00fb6f6144	karolinska	0	0+0
6	003046e27c8ead3e3db155780dc5498e	karolinska	1	3+3
7	0032bfa835ce0f43a92ae0bbab6871cb	karolinska	1	3+3
8	003a91841da04a5a31f808fb5c21538a	karolinska	1	3+3
9	003d4dd6bd61221ebc0bfb9350db333f	karolinska	1	3+3

Fig. 6. First 10 Lines of the Dataframe of the Training Data

weights from the "imagenet" dataset, the model leverages a wealth of learned visual features. Its deep architecture, with residual connections, allows for the extraction of intricate patterns and improved representation learning. Dropout regularization helps prevent overfitting, while the inclusion of a fully connected layer with ReLU activation enables nonlinear transformations. The RMSprop optimizer and binary cross-entropy loss function ensure efficient training and accurate predictions. Evaluation metrics such as binary accuracy and AUC provide a comprehensive assessment of the model's performance. By saving the trained model and visualizing the training history, researchers can effectively communicate the progress and capabilities of the model.

2.6 VGG19

The VGG19 model implementation in the study leverages the pre-trained weights from the "imagenet" dataset, allowing for the extraction of meaningful features from prostate cancer images. The architecture consists of multiple convolutional layers with 3×3 filters, maximizing feature extraction capabilities. Dropout regularization is applied to reduce overfitting, while dense layers with ReLU activation contribute to the classification process. The model is trained using the binary cross-entropy loss function and RMSprop optimizer, and key evaluation metrics such as accuracy and AUC are monitored. By visualizing the training history, the model's performance can be assessed, providing valuable insights for further research and application in prostate cancer detection.

2.7 InceptionV3

The InceptionV3 model, based on the Inception architecture, is employed in this research. Pre-trained weights from the "imagenet" dataset provide a strong initialization for the model. The architecture utilizes Inception modules with parallel convolutional layers of varying sizes to capture intricate features at different scales. Dropout regularization is incorporated to prevent overfitting, and a fully connected layer with 32 units and ReLU activation is added.

The final output layer utilizes the softmax activation function to produce class probabilities. The model is compiled with binary cross-entropy loss and the RMSprop optimizer, while evaluation metrics such as accuracy and AUC provide insight into its performance. During training, the fit-generator function is employed, allowing for efficient batch processing of the training and validation data. Callbacks for early stopping, model checkpointing, and learning rate scheduling contribute to better training results.

The trained model is saved, and the training history is visualized using plots that depict the loss, AUC, and accuracy curves for both the training and validation sets. This comprehensive implementation of the InceptionV3 model demonstrates its potential for accurate prostate cancer classification and can be a valuable contribution to research studies in the field.

3 Results

With the implementation of the models, three graphs were obtained representing the train and validation Loss, train and validation AUC, and train and validation Accuracy for each of them.

3.1 Performance of Resnet-50 Model

With the implementation of the ResNet-50 model, we were able to characterize its behavior by analyzing the curves of the graphs. The Train Loss curve behaves in a manner considered normal, indicating that the model is learning and reducing its loss during the training process. However, the Validation Loss curve suggests the presence of underfitting during model training. This indicates that the model was unable to sufficiently capture the complexity of the data and may not generalize well to unseen examples.

In the AUC graph, both the validation and training curves show an increasing trend as the model is trained. This indicates that the model is capturing and learning relevant discriminative patterns from the training data.

Analyzing the Accuracy graph, the train accuracy curve shows a positive behavior, indicating a good fit of the model to the training data. However, for a more detailed evaluation, it is necessary to consider the validation accuracy curve. In this case, it suggests that the model initially struggled to generalize well to the validation data but eventually improved its performance by adjusting to the patterns present in that data (Fig. 7).

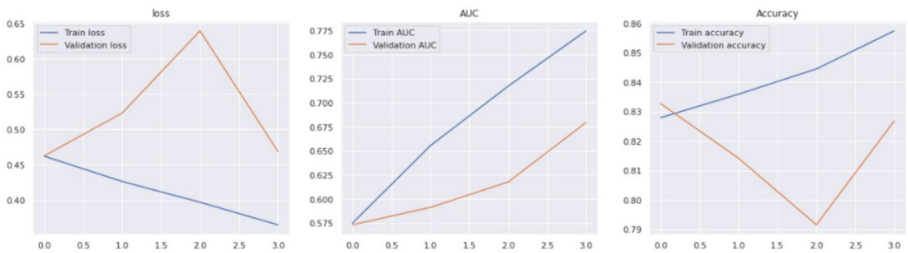

Fig. 7. Graphical Representation of the Metrics for the ResNet-50 model

3.2 Performance of VGG19 Model

With the implementation of the VGG19 model, we were able to analyze its performance using the same metrics. By examining the loss representation, we can observe that the Train Loss curve decreases during training, indicating that the error between the model's predictions and the actual values is decreasing. The

Validation Loss curve also decreases, indicating that the model's performance improves as it learns to make more accurate predictions on unseen data.

In the AUC graph, we can see that both the training and validation curves increase during training, indicating that the model's performance improves on both the training and validation data as it progresses. This means that the model becomes more capable of correctly distinguish between the classes of interest.

In the accuracy graph, we observe a similar behavior to the AUC curves. Both the training and validation curves increase during training, which can be the result of various improvements in the model, such as proper parameter tuning, appropriate feature selection, or regularization. The increasing curves indicate that the model becomes more skilled at making correct predictions and performs good generalization to new data (Fig. 8).

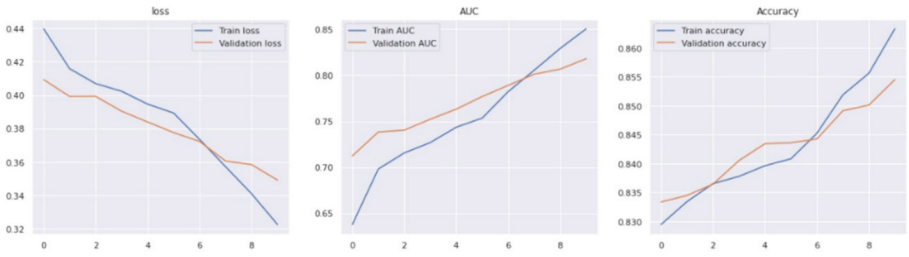

Fig. 8. Graphical Representation of the Metrics for the VGG19 model

3.3 Performance of InceptionV3 Model

With the implementation of the InceptionV3 model, the following results were obtained: in the loss representation graph, a decreasing trend in the training curve and a parabolic shape in the validation curve can be observed. The decreasing training curve over time is expected during this process, as it indicates that the model is progressively adjusting to the training data and reducing prediction errors. The parabolic shape in the validation curve may indicate that the model is reaching a point where it is overfitting to the training data, resulting in higher loss on the validation data.

In the AUC representation, an increase is observed in both the training and validation curves. This can be considered a positive sign, indicating that the model is learning to capture discriminative patterns in the training data and is able to generalize this learning to the validation data.

In the Accuracy representation, the training curve increases over time, while the validation curve initially decreases and then increases, with a slight decrease towards the end. The increasing training curve signifies that the model is fitting more closely to the training data. The shape of the validation curve indicates a common training pattern. The initial decrease suggests that the model is still

learning to generalize the patterns from the training data to unseen data. The subsequent increase may represent natural fluctuations in the validation data or the introduction of more challenging training steps. Finally, the slight decrease observed could be a sign of overfitting, indicating that the model may be becoming too specific to the training data and does not generalize well to unseen data (Fig. 9).

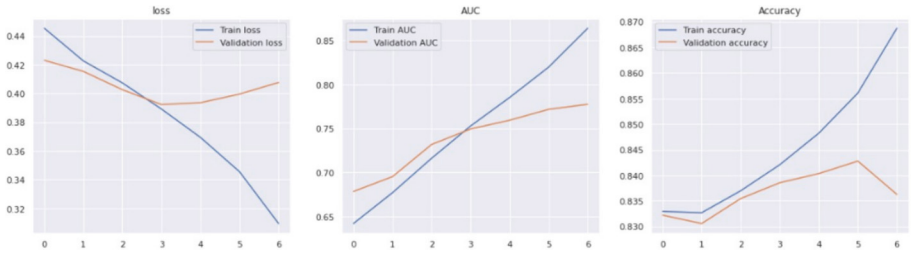

Fig. 9. Graphical Representation of the Metrics for the InceptionV3 model

3.4 Comparison Between the Data and the Predictions

Finally, a comparison was made between the ISUP grade of the original data and the predicted data for the first 40 rows of the validation file. This comparison revealed that the ISUP grade in the original data and the predicted data coincided, as shown in the Fig. 10.

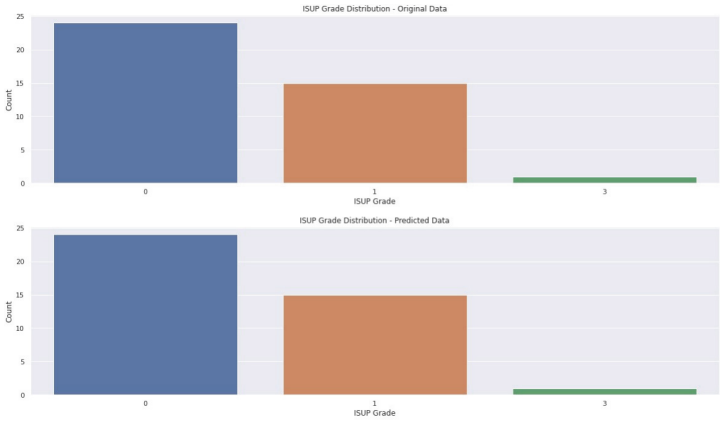

Fig. 10. Comparison between the data and the predictions

4 Discussion

The results obtained with the implementation of the ResNet-50 model show a mixed behavior in terms of the model's performance. On one hand, the Train Loss curve follows an expected trend, gradually decreasing during training, indicating that the model is learning and reducing its error as it is exposed to more data. However, the Validation Loss curve suggests the presence of underfitting, indicating that the model was unable to adequately capture the complexity of the data and may not generalize well to unseen examples. In the AUC graph, both the validation and training curves show an increasing trend as the model is trained. This is a positive indication, suggesting that the model is capturing and learning relevant discriminative patterns from the training data. Analyzing the Accuracy graph, the train accuracy curve shows a positive behavior, indicating a good fit of the model to the training data. However, for a more detailed evaluation, it is necessary to consider the validation accuracy curve. In this case, it suggests that the model initially struggled to generalize well to the validation data but eventually improved its performance by adjusting to the patterns present in that data.

The results obtained with the implementation of the VGG19 model demonstrate positive performance across multiple metrics. In terms of loss, both the Train Loss curve and Validation Loss curve consistently decrease during training. This indicates that the model's predictions are becoming more accurate, as the error between the predicted values and the actual values decreases. This improvement suggests that the model is effectively learning to make accurate predictions on unseen data. The AUC graph further confirms the model's performance improvement. Both the training and validation curves exhibit an increasing trend as the model is trained. This indicates that the model is becoming more proficient at distinguishing between the classes of interest, as it captures and learns relevant discriminative patterns from the training data. The increasing curves demonstrate the model's ability to generalize its learning to new and unseen data. Similar behavior is observed in the accuracy graph. Both the training and validation curves consistently increase during training, indicating that the model's predictive accuracy improves over time. This improvement can be attributed to various factors, such as appropriate parameter tuning, suitable feature selection, or regularization techniques. The increasing curves indicate that the model becomes more adept at making correct predictions and demonstrates good generalization performance on new data.

The implementation of the InceptionV3 model yielded specific results across different metrics. In terms of loss representation, the training curve shows a decreasing trend, indicating that the model is progressively adjusting to the training data and reducing prediction errors. This behavior is expected during the training process. However, the validation curve exhibits a parabolic shape, suggesting that the model may be reaching a point of overfitting. Overfitting occurs when the model becomes too specific to the training data, resulting in higher loss on the validation data. This indicates that the model may struggle to generalize well to unseen data. Moving on to the AUC representation, both the training and validation curves display an increase. This is a positive indication,

demonstrating that the model is successfully learning and capturing discriminative patterns from the training data. The ability to generalize this learning to the validation data suggests that the model is effectively identifying and distinguishing between different classes. Examining the accuracy representation, the training curve consistently increases over time. This signifies that the model is fitting more closely to the training data and improving its performance. On the other hand, the validation curve initially decreases and then gradually increases, with a slight decrease observed towards the end. The initial decrease suggests that the model is still in the process of learning to generalize the patterns from the training data to unseen data. The subsequent increase could be attributed to natural fluctuations in the validation data or the introduction of more challenging training steps. However, the slight decrease towards the end may indicate overfitting, implying that the model is becoming too specific to the training data and struggling to generalize well to unseen data.

The alignment observed in the comparison between the ISUP grade of the original and predicted data for the first 40 rows of the validation file indicates that the trained models were able to effectively learn the patterns present in the training data and generalize well to the same data (Fig. 11).

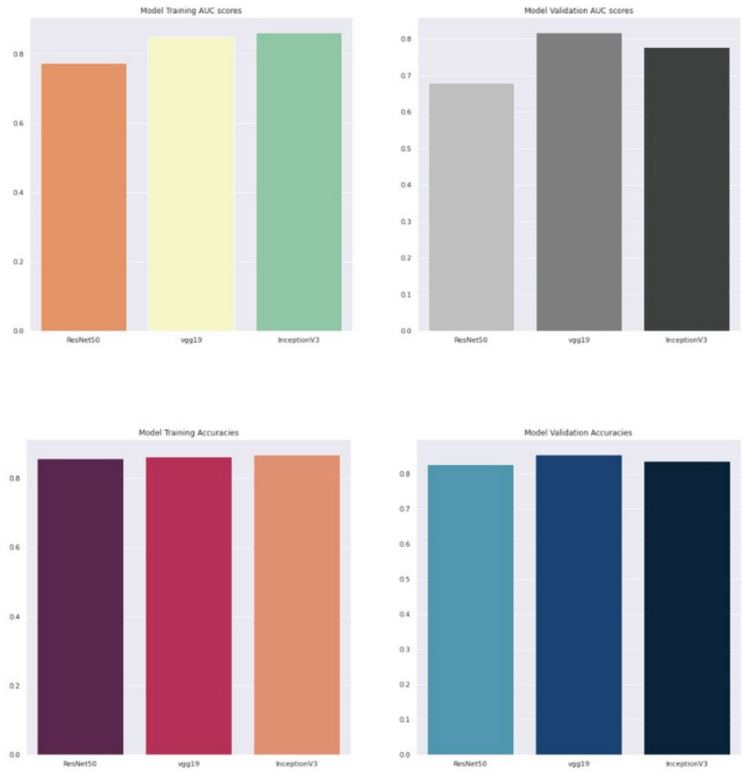

Fig. 11. Graphical representation of Training and Validation for each model

5 Conclusions

Based on these results, it can be concluded that this approach is potentially useful for determining the ISUP grade when provided with training data and masks that have annotated Gleason scores. The ResNet-50, VGG19, and InceptionV3 models demonstrated varying performances in capturing relevant patterns and generalizing to unseen data.

While the ResNet-50 model showed mixed behavior with signs of underfitting, it still captured relevant patterns from the training data. The VGG19 model exhibited positive performance across multiple metrics, indicating accurate predictions and good generalization capabilities. The InceptionV3 model showed potential but struggled with overfitting and generalization to unseen data.

The alignment between the ISUP grade of the original and predicted data for the first 40 rows of the validation file further suggests that the trained models effectively learned the patterns and were able to generalize well to the same data. This indicates that the models have the potential to accurately predict the ISUP grade based on the provided training data and masks with annotated Gleason scores.

However, further evaluation and testing on additional datasets are necessary to validate the performance of these models and determine their reliability in real-world scenarios.

References

1. Epstein, J., Amin, M., Fine, S., Algaba, F., Aron, M.: The 2019 genitourinary pathology society (GUPS) white paper on contemporary grading of prostate cancer. Arch. Pathol. Lab. Med. **145**, 461–493 (2020)
2. Leenders, G., Verhoef, E., Hollemans, E.: Prostate cancer growth patterns beyond the Gleason score: entering a new era of comprehensive tumour grading. Histopathology **77**, 850–861 (2020)
3. Bulten, W., et al.: Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. Mod. Pathol. **34**, 660–671 (2021)
4. Bulten, W., et al.: Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat. Med. **28**, 154–163 (2022)
5. Kartasalo, K., et al.: Artificial intelligence for diagnosis and Gleason grading of prostate cancer in biopsies-current status and next steps. Eur. Urol. Focus **7**, 687–691 (2021)
6. Egevad, L., et al.: Identification of areas of grading difficulties in prostate cancer and comparison with artificial intelligence assisted grading. Virchows Archiv **477**, 777–786 (2020)
7. Marginean, F., et al.: An artificial intelligence-based support tool for automation and standardisation of Gleason grading in prostate biopsies. Eur. Urol. Focus **7**, 995–1001 (2021)
8. Pantanowitz, L., et al.: An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. Lancet Digit. Health **2**, e407–e416 (2020)