Machine Learning Methods for Prostate Cancer Prediction using Magnetic Resonance Imaging and Clinical Data

Singh Akal Ustat[1], Thiramdas Vikas Balram[1], Baynes Anna[1], Bang Tran[1]

[1] Department of Engineering & Computer Science, California State University 6000 J St, Sacramento, CA 95819

Abstract

Prostate cancer is one of the most common cancers in males and is difficult to diagnose in its early stages. Traditional methods have attempted to predict prostate cancer from clinical data, but lack of ability to incorporate image to enhance the predictive outcomes. To address this issue, we develop an analysis pipeline that integrates clinical and medical imaging data to efficiently predict prostate cancer using multiple machine learning approaches. In this paper, we demonstrate a high performance of our analysis pipeline by using Magnetic Resonance Imaging, clinical data and advanced pre-processing technique to overcome class imbalance challenges. In the task of analyzing prostate cancer datasets from Challemon Challenge, our analysis pipeline delivers a high accuracy score of 86.9% in prediction using deep neural network architecture, showing strong superiority over traditional methods. This research is not only promising in providing improvement in patient outcomes but also optimizing use of healthcare resources.

1. Introduction

Prostate cancer remains among the most common cancers diagnosed in males worldwide. It is projected to account for approximately 1.4 million new cases for the year 2020 and continues to rise as a result of demographic and lifestyle factors [1, 2]. Detection and proper risk stratification are some of the important components of effective management of prostate cancer because these will dictate clinical decision-making and consequently determine the outcome of treatment. Traditional diagnostic methods like the Prostate Specific Antigen (PSA) test and Digital Rectal Examination (DRE) suffer from low specificity and sensitivity, leading to over-diagnosis, unnecessary biopsies, and missed aggressive cancers [3, 4], highlighting the need for a more accurate, scalable, and non-invasive alternative.

Medical imaging modalities including MRI and TRUS, or Transrectal Ultrasound, have enhanced prostate abnormality visualization and have allowed for the detection of early-stage disease [5–7]. But these imaging technologies are costly, requiring expert interpretation. They are not scalable nor accessible in clinical settings because of their resource intensity. In addition, with the inclusion of imaging data along with clinical variables, such as PSA levels, patient demographics, and family history, the computation also becomes quite challenging, necessitating more sophisticated analytical methods [8, 9]. ML is a method of AI, which can successfully solve this challenge. Since ML uses complex, high-dimensional data, it potentially detects patterns and relationships in data that more traditional statistical methods could not uncover [10, 11].

Deep learning methods in ML, especially CNNs, have shown exceptional results in medical image analysis [12, 13]. CNNs are suitable for the task like tumor detection, segmentation and feature extraction from imaging data, so especially suitable for prostate cancer diagnosis where there is the issue of great heterogeneity of data [10, 12]. Use of combination of CNN with traditional classifier such as KNN, Naive Bayes presents an effective methodology for measuring the effectiveness of different techniques used for estimation of prognosis of prostate cancer [14,15].

This study bridges clinical relevance and computational innovation by developing an integrated pipeline for prostate cancer prediction. It combines clinical data and image analysis while addressing class imbalance and data heterogeneity. Traditional machine learning and deep learning models are compared to identify the most effective approach for severity prediction. Additionally, CNN architectural design strategies are explored to reduce overfitting and enhance generalization [16, 17].

2. Methods

Figure 1 shows the overall prediction pipeline of our approach. The pipeline accepts clinical features and high-dimensional 3D MRI scans as inputs. The inputs data go through a pre-processing step, involving resizing the MRI scans and using the Synthetic Minority Over-sampling Technique (SMOTE) to balance the class distributions.

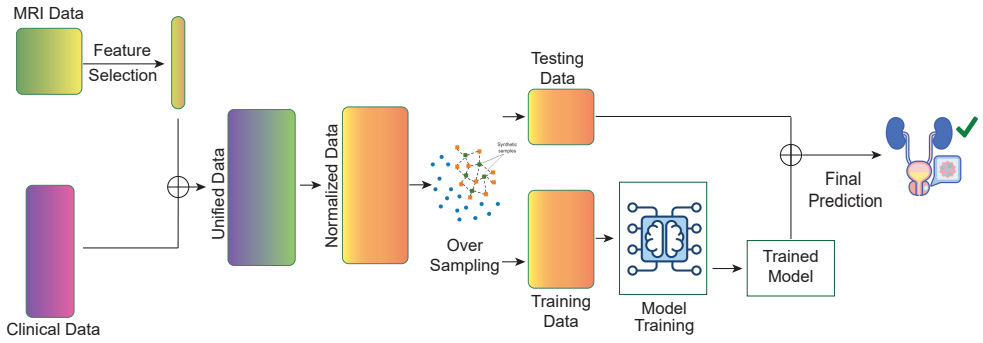

Figure 1. The overview of the prostate cancer prediction pipeline. This pipeline uses MRI and clinical data as primary inputs. After feature selection, data is normalized and over-sampled using synthetic data generation to address class imbalance. The dataset is then split into training and testing sets, where the model is trained and evaluated. The trained model makes the final prediction, ensuring a robust and reliable diagnostic process.

Next, we split the processed data into training and testing sets. Then we train and evaluate multiple machine learning models on the training set. Finally, we used the best model obtained from the training phase to make predictions on the testing set.

2.1. Data Processing

This study uses clinical and with 3D MRI scans as the primary inputs to diagnose prostate cancer. MRI images are obtained in a sequential thin sections called a slice. Each slice represents a cross-sectional view of the body in a specific plane. The thickness and spacing of these slices may vary based on the imaging protocol and desired resolution. The number of slices are varied per experiment. We choose to use 40 first slices for our analysis. From selected slices, we encode the inputs feature using 3D Convolutional Neural Network. Finally, we get a flatten MRI data in a 1D array [18, 19], converting the volumetric information into a format suitable for machine learning models. This transformation helps reduce dimensionality while preserving essential patterns for analysis. Then, we merge the flatten data obtained from the MRI with clinical information to consolidate them into a unified dateset.

We also notice that the dataset has an extreme class imbalance because the majority class (low-severity) comprised 71% of the samples, while the minority class (high-severity) comprised only 29%. We apply SMOTE [20]to balance the class distribution. Figure 2 show the side-by-side of class distribution before and after applying SMOTE. Finally, we use 70% of data samples for training and the rest is used for testing.

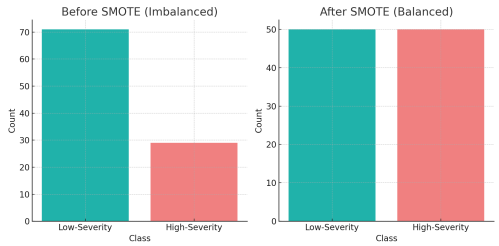

Figure 2. Impact of the Synthetic Minority Over-sampling Technique on balancing class distributions in the dataset. The left panel shows the original dataset, where the Low-Severity class (blue) comprised 71% of the samples, while the High-Severity class (red) made up only 29%, leading to potential model bias. The right panel depicts the dataset after SMOTE, where both classes are balanced at 50%.

2.2. Classical Models

After data preprocessing step, we use the obtained data to train multiple machine learning models. Here, we aim at using models that belong classical and deep learning approaches.

The first classical method we use is K-Nearest Neighbors (KNN) [21] which is a simple yet powerful algorithm that classifies instances based on the majority class of their k-nearest neighbors. The KNN decision rule is defined as:

$$\hat{y} = \arg\max_{y} P(y|\mathbf{x}) = \arg\max_{\mathbf{y}} \frac{\mathbf{1}}{\mathbf{k}} \sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{I}(\mathbf{y} = \mathbf{y_i}) \quad (1)$$

where $\hat{y}$ is the predicted class, $\mathbf{x}$ are the input features, y_i are the labels of the k-nearest neighbors, and $I(\cdot)$ is an indicator function, which outputs one if the condition holds.

Hyperparameter tuning is performed via grid search to optimize k, minimizing prediction bias and variance. In this paper, we use $k = 10$ and Euclidean distance for training.

Second, we use Naïve Bayes [22] classifier which is probabilistic model based on Bayes' Theorem. It is assumed that features are conditionally independent given the class labels. The posterior probability of each class is computed as follows:

$$P(y|\mathbf{x}) = \frac{\mathbf{P}(\mathbf{x}|\mathbf{y}) \cdot \mathbf{P}(\mathbf{y})}{\mathbf{P}(\mathbf{x})} \tag{2}$$

where $P(y|\mathbf{x})$ is the posterior probability of class y given the input $\mathbf{x}$, $P(\mathbf{x}|\mathbf{y})$ is the likelihood of $\mathbf{x}$ given y, $P(y)$ is the prior probability of y, and $P(\mathbf{x})$ is the evidence. Despite relying on many independent assumptions, Naïve Bayes works particularly well on large training sets where the independence of features approximately holds Euclidean distance. In this paper, we use a smoothing parameter to avoid zero-probability issues, improve generalization, and prevent overfitting.

2.3. Deep Learning Models

In the deep learning approach, we use CNN [23]model which is particularly suited for extracting spatial features from the structure of images such as MRI scans, even distorted, translated, or rotated images. The CNN architecture consists of convolutional layers followed by fully connected layers to capture these hierarchical patterns. Each convolutional layer filters the processed input data to extract relevant features such as textures and spatial relationships. Then, the fully connected layers learn how to combine these features to make predictions about the severity of the cancer. Here, we evaluate two different CNN model's architecture: (i) Single-Layer CNN Architecture and (ii) Multi-layer CNN Architecture.

Single-layer CNN used in this study applies convolutions over data using filters or kernels to pick up relevant features in some input region. This operation is defined as:

$$Z_{i,j,k} = \sum_{m,n} X_{i+m,j+n} \cdot W_{m,n,k} + b_k \tag{3}$$

where X is the input image, W represents the filter weights, and b_k is the bias term for the *k-th* filter, and Z is the output feature map. Here, we use Rectified Linear Unit (ReLU) as activation function. ReLU provides nonlinearity through replacement of all negative values with zero using the following formula:

$$f(x) = \max(0, x) \tag{4}$$

To accelerate the training process without compromising model's performance, we use dropout regularization

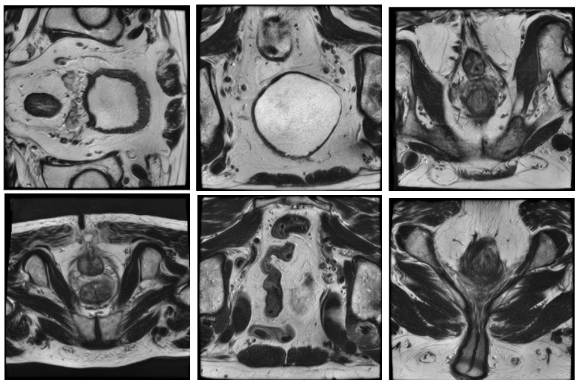

Figure 3. Visualization of the processed MRI obtained from six randomly selected patients . Each image shows one volume slide sampled from multiple images taken for each patient. Images highlight the spatial resolution and intensity variations.

[24] technique. Here, a fraction of p neurons are randomly deactivated during training to prevent over-fitting:

$$h_i = \begin{cases} x_i, & \text{with probability } 1-p \\ 0, & \text{with probability } p \end{cases} \tag{5}$$

Fully connected layer is added after the CNN layer. The fully connected layer classifies the data as follows:

$$\hat{y} = \sigma(W_{fc} \cdot h + b_{fc}) \tag{6}$$

where W_{fc} and b_{fc} are the weights and biases of the fully connected layer, h is the flattened feature vector, and σ represents the sigmoid activation function.

Finally, we use multi-layer CNN architecture as a second deep learning model. A multi-layer CNN increases the number of convolutional layers to extract increasingly abstract features. As a result, the network captures more complex features. Max-pooling layers further downsample and reduce the data's dimensionality, minimizing the model's computational requirements while retaining important features. Additionally, we batch normalization to stabilize training by normalizing the inputs to each layer, improving convergence alongside performance.

3. Result

In this section, we assess the performance of multiple machine learning methods in the ability of accurate classification of prostate cancer.

3.1. Data Preparation

This section presents the process of preparing data for our analyses. We download 3D MRI image and clinical prostate cancers data from EU-funded CHAIMELEON

Project. The repository contains data from 295 patient records with clinical features and 3D MRI scans. The clinical data was store in JSON format where patient ID, age, and Prostate-Specific Antigen levels are available. The ground truth information of 'High' and 'Low' severity classes are also available and we only use ground truth for training.

Figure 3 shows a representative slice randomly taken from multiple volumes and patients in 3D MRI dataset. Visualizing individual slices helps to understand the underlying characteristics of MRI data, such as spatial resolution, intensity variations, and potential imaging artifacts as well as the quality and integrity of the data for downstream pre-processing and analysis. We make the spatial resolution consistent by resizing them to a standard dimension of $256 \times 256 \times 40$ voxels through interpolation techniques[25], reducing variation due to different scanning protocols.

3.2. Model performance

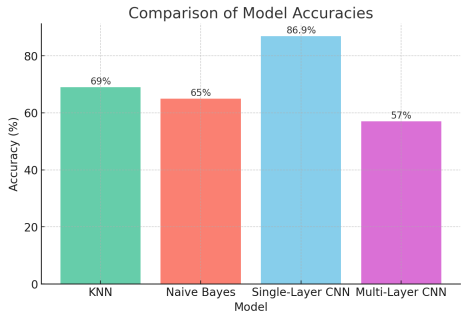

Figure 4. The accuracy of four classification models, with each represented in distinct pastel colors. The Single-Layer CNN achieved the highest accuracy (86.9%), while the Multi-Layer CNN exhibited overfitting, resulting in the lowest accuracy (57%).

Figure 4 shows the overall performance of all the assessed models. Here, we measure the ability to accurately predicted patients into the respected classes. Here, KNN classifier achieves 69% accuracy, with a moderate precision of 70% and recall of 60%. This indicates a limited capability for capturing complex patterns within MRI data. The reliance on proximity metrics such as Euclidean distance failed to capture the non-linear relationships and spatial complexities inherent to MRI scans. Moreover, hyperparameter tuning may not have sufficiently accounted for these, either. Similarly, Naïve Bayes classifier achieves 65% with precision of 71% and recall of 67%., the model performs relatively well for low-severity cases. This is consistent with our understanding of Naïve Bayes because it relies on an assumption of feature independence. However, it struggled with MRI images, where these assumptions break down; spatial dependencies abound, and complex feature interactions are inherent to MRIs.

The challenges experienced with KNN and Naïve Bayes make it evident that high-dimensional MRI images require advanced deep learning models. The single-layer CNN achieved 86.9% accuracy on the test set, showing promising results. On the other hand, although more complex, the multi-layer CNN suffered from severe over-fitting, resulting in an accuracy of 57%. In summary, complex architectures tend to over-fit more, while simpler architectures fail to capture complex patterns. This highlights the need for carefully selecting model architecture and regularization techniques to balance performance and generalization.

4. Conclusion

In this paper, we proposed a machine-learning pipeline for prostate cancer detection using classical machine learning and deep learning data analysis. These techniques use low-dimensional data (e.g., patient demographics and PSA) and high-dimensional data such as MRI scans. A further comparison reveals how classical algorithms are better suited to low-dimensional data, failing to achieve meaningful accuracy for the MRI scans. On the other hand, deep learning models, which achieve a balance between complexity and accuracy, perform better at cancer detection using image data. In the the future we plan to integrate denoising techniques [18, 26] and sub-typing methods [27] to improve the predictive outcomes and sub-disease group discovery.

Acknowledgments

This work was partially supported by Carole McNamee Student/Faculty Research Endowment, CSUS Computer Science Department, CSUS Research Enhanced Support Grants. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer Journal for Clinicians 2018; 68(6):394–424.

[2] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA a Cancer Journal for Clinicians 2019;69(1):7–34.

[3] Catalona WJ, Smith DS, Ratliff TL, Dodds KM, Coplen DE, Yuan JJ, Petros JA, Andriole GL. Measurement of prostate-specific antigen in serum as a screening test for

prostate cancer. New England Journal of Medicine 1991; 324(17):1156–1161.

[4] Schröder FH, Kruger AB, Rietbergen J, Kranse R, Maas Pvd, Beemsterboer P, Hoedemaeker R. Evaluation of the digital rectal examination as a screening test for prostate cancer. Journal of the National Cancer Institute 1998; 90(23):1817–1823.

[5] Panzone J, Byler T, Bratslavsky G, Goldberg H. Transrectal ultrasound in prostate cancer: current utilization, integration with mpmri, hifu and other emerging applications. Cancer Management and Research 2022;1209–1228.

[6] Lee F, McHugh T, Solomon M, Dorr R, Siders D, Kirscht J, Christensen L, Mitchell A. Transrectal ultrasound, digital rectal examination, and prostate-specific antigen: preliminary results of an early detection program for prostate cancer. Scandinavian Journal of Urology and Nephrology Supplementum 1991;137:101–105.

[7] Hambrock T, Hoeks C, Hulsbergen-Van De Kaa C, Scheenen T, Fütterer J, Bouwense S, Van Oort I, Schröder F, Huisman H, Barentsz J. Prospective assessment of prostate cancer aggressiveness using 3-t diffusion-weighted magnetic resonance imaging–guided biopsies versus a systematic 10-core transrectal ultrasound prostate biopsy cohort. European urology 2012;61(1):177–184.

[8] Ahmed HU, Bosaily AES, Brown LC, Gabe R, Kaplan R, Parmar MK, Collaco-Moraes Y, Ward K, Hindley RG, Freeman A, et al. Diagnostic accuracy of multi-parametric mri and trus biopsy in prostate cancer (promis): a paired validating confirmatory study. The Lancet 2017;389(10071):815–822.

[9] Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, Margolis D, Schnall MD, Shtern F, Tempany CM, et al. Pi-rads prostate imaging–reporting and data system: 2015, version 2. European Urology 2016;69(1):16–40.

[10] Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H. Computer-aided detection of prostate cancer in mri. IEEE Transactions on Medical Imaging 2014;33(5):1083–1092.

[11] Tătaru OS, Vartolomei MD, Rassweiler JJ, Virgil O, Lucarelli G, Porpiglia F, Amparore D, Manfredi M, Carrieri G, Falagario U, Terracciano D, Cobelli Od, Busetto, Maria G, Giudice FD, Ferro M. Artificial intelligence and machine learning in prostate cancer patient management—current trends and future perspectives. Diagnostics 2021;11(2):354.

[12] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annual Review of Biomedical Engineering 2017; 19(1):221–248.

[13] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JA, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Medical Image Analysis 2017;42:60–88.

[14] Yoo S, Gujrathi I, Haider MA, Khalvati F. Prostate cancer detection using deep convolutional neural networks. Scientific Reports 2019;9(1):1–10.

[15] Pellicer-Valero OJ, Jiménez JLM, Gonzalez-Perez V, Ramón-Borja JLC, García IM, Benito MB, et al. Deep learning for fully automatic detection, segmentation, and gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. Scientific Reports 2021; 11:1–12.

[16] Cao R, Bajgiran AM, Mirak SA, Shakeri S, Zhong X, Enzmann DR, Raman SS, Sung K. Joint prostate cancer detection and gleason score prediction in mpmri via focalnet. IEEE Transactions on Medical Imaging 2019;38(11):2496–2506.

[17] Hussein S, Cao R, Sung K, Raman SS, Patel P, Reiter RE, et al. Deep learning-based pipeline for prostate cancer grading in multiparametric mri with clinical integration. Medical Image Analysis 2022;75:102267.

[18] Tran B, Tran D, Nguyen H, Ro S, Nguyen T. sccan: single-cell clustering using autoencoder and network fusion. Scientific Reports 2022;12(1):10267.

[19] Tran D, Nguyen H, Tran B, La Vecchia C, Luu HN, Nguyen T. Fast and precise single-cell data analysis using a hierarchical autoencoder. Nature Communications 2021; 12(1):1029.

[20] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 2002;16:321–357.

[21] Guo G, Wang H, Bell D, Bi Y, Greer K. Knn model-based approach in classification. In Meersman R, Tari Z, Schmidt DC (eds.), On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, volume 2888 of Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. ISBN 978-3-540-39964-3, 2003; 986–996.

[22] McCallum A, Nigam K. A comparison of event models for naive bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization. 1998; 41–48. URL https://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf.

[23] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998;86(11):2278–2324. ISSN 0018-9219.

[24] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 2014;15(1):1929–1958.

[25] Thévenaz P, Blu T, Unser M. Interpolation revisited. IEEE Transactions on Medical Imaging 2000;19(7):739–758.

[26] Tran B, Nguyen Q, Shrestha S, Nguyen T. scids: Single-cell imputation by combining deep autoencoder neural networks and subspace regression. In 2021 13th International Conference on Knowledge and Systems Engineering (KSE). IEEE, 2021; 1–8.

[27] Nguyen H, Tran D, Tran B, Roy M, Cassell A, Dascalu S, Draghici S, Nguyen T. Smrt: Randomized data transformation for cancer subtyping and big data analysis. Frontiers in Oncology 2021;11:725133.

Address for correspondence:

Bang Tran
Department of Engineering & Computer Science, California State University 6000 J St, Sacramento, CA 95819
s.tran@csu.edu