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

Problem Statement

- Prostate Cancer is among the most common cancer diagnosed in males worldwide.
  - Approximately 1.4 million new cases in 2020 and expected to increase further.
- Traditional Diagnostic methods like PSA testing and DRE suffer from low specificity and sensitivity
  - Leads to overdiagnosis, unnecessary biopsies, and false negatives.
- MRI and TRUS technologies, although useful, are not scalable due to cost, accessibility, and need for expertise.

Proposed Solution

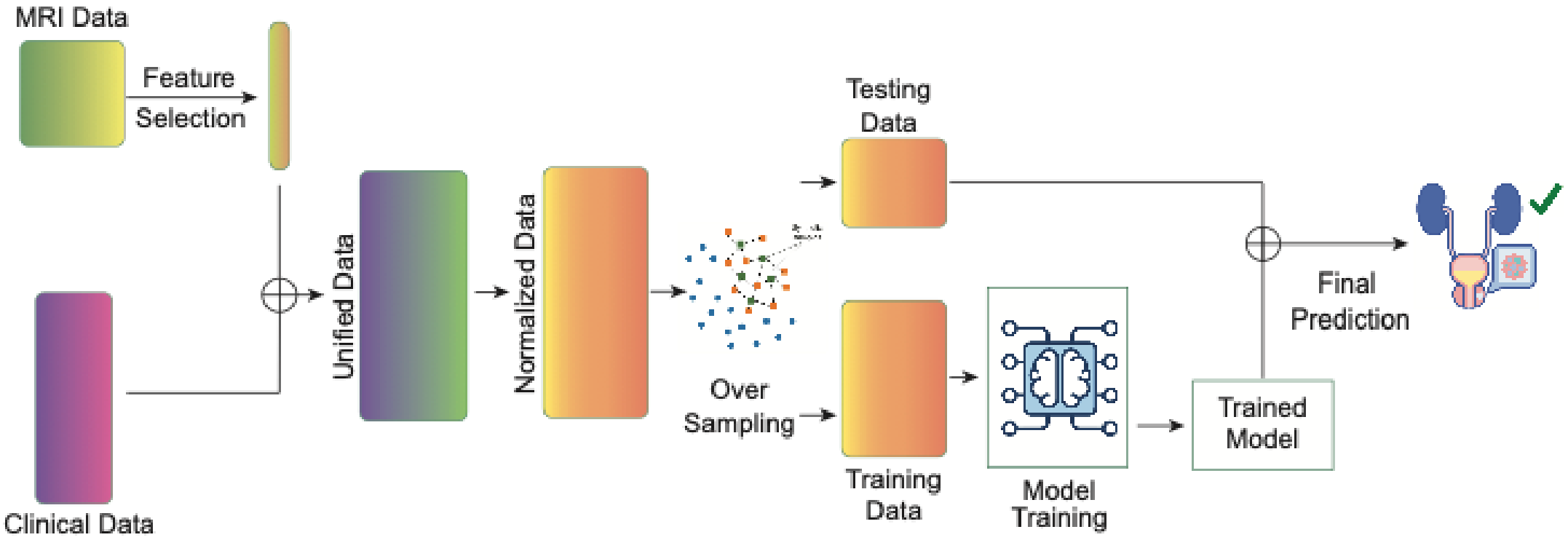
We can use deep learning methods such as CNNs to address these drawbacks.

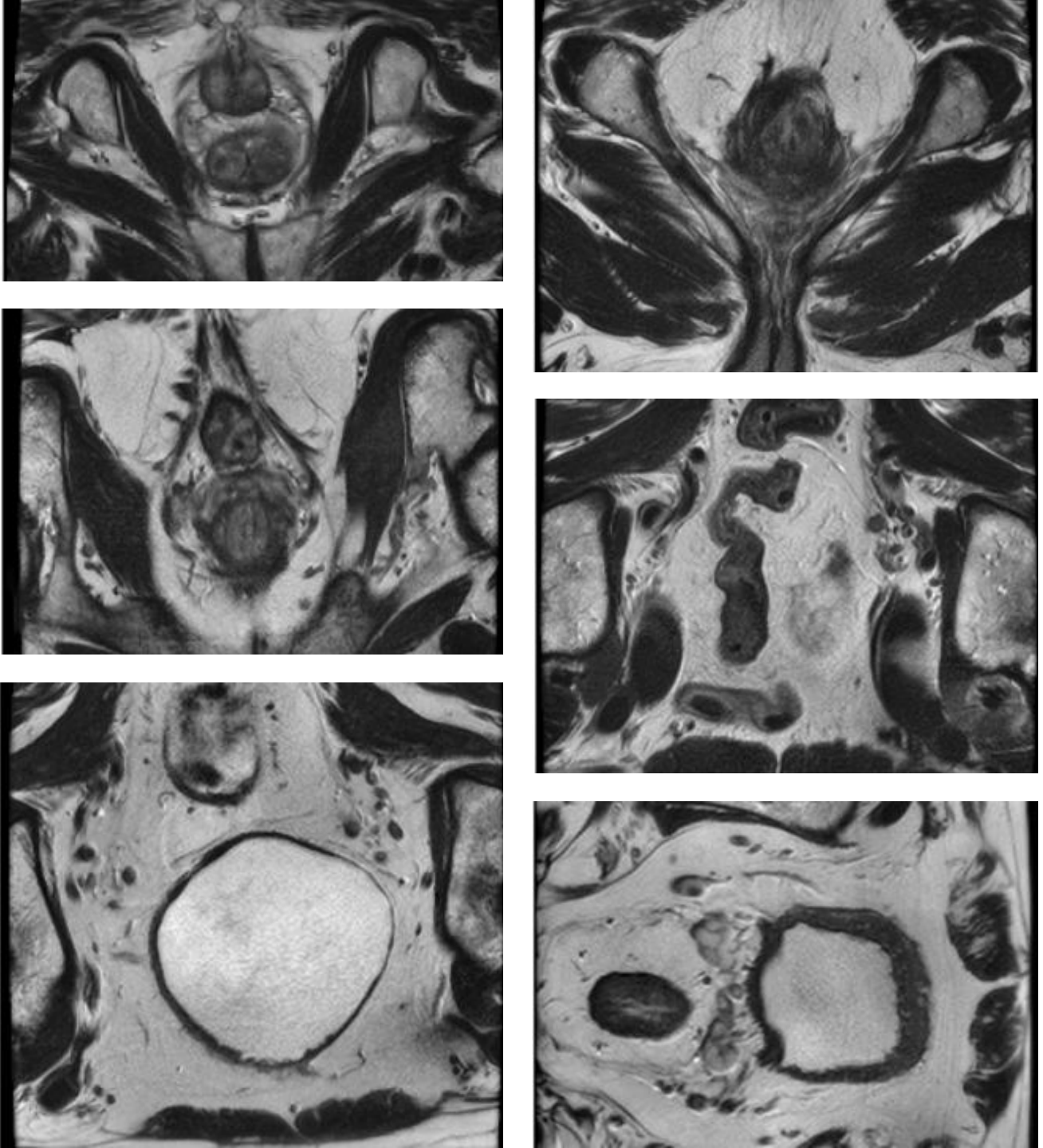
CNNs have exceptional results in medical imaging analysis, since they can detect relationships in complex, high-dimensional data.

Furthermore, we can measure CNN effectiveness by comparing results with traditional classifiers like KNN and Naïve Bayes.

Methodology

- The prediction pipeline outlined by our approach is as follows:
  - 3D MRI scans are pre-processed, which includes resizing and SMOTE to balance class distributions.
  - Data is split into training and testing sets.
  - Multiple machine learning models are trained and evaluated on the training set.
  - Finally, the best model is used for prediction on the testing set.



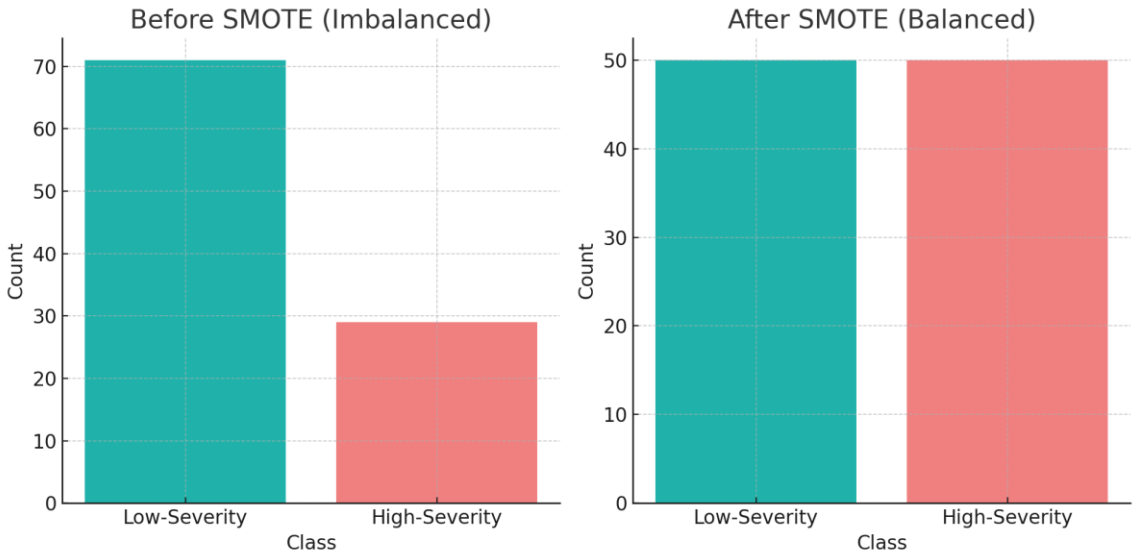


Step 1: Data Processing

- The MRI images used in this study are obtained in thin sequential sections called slices.
  - Cross-sectional views of the body in a specific plane.
  - Our analysis uses 40 slices.
- Inputs are encoded using a 3D CNN and then flattened into a 1D array.
  - Reduces dimensionality while preserving essential patterns
- The MRI data is then merged with clinical information to create a unified dataset.

Step 1: Data Processing (Cont'd.)

- 3D MRI data and clinical data was obtained from the EU-funded CHAIMELEON Project.
- The data set, however, has an extreme class imbalance.
  - Low-severity samples far outnumbers high-severity samples
- SMOTE is applied to balance the distribution using synthetic data.



Step 2: Usage of Classical Models

KNN is first preformed in the data.

Hyperparameter tuning is performed via grid search

- k = 10 is used along with Euclidean distance for training.

Naïve Bayes Classifier is also used.

This classifier works well on larger training sets where the independence of features approximately holds Euclidean distance.

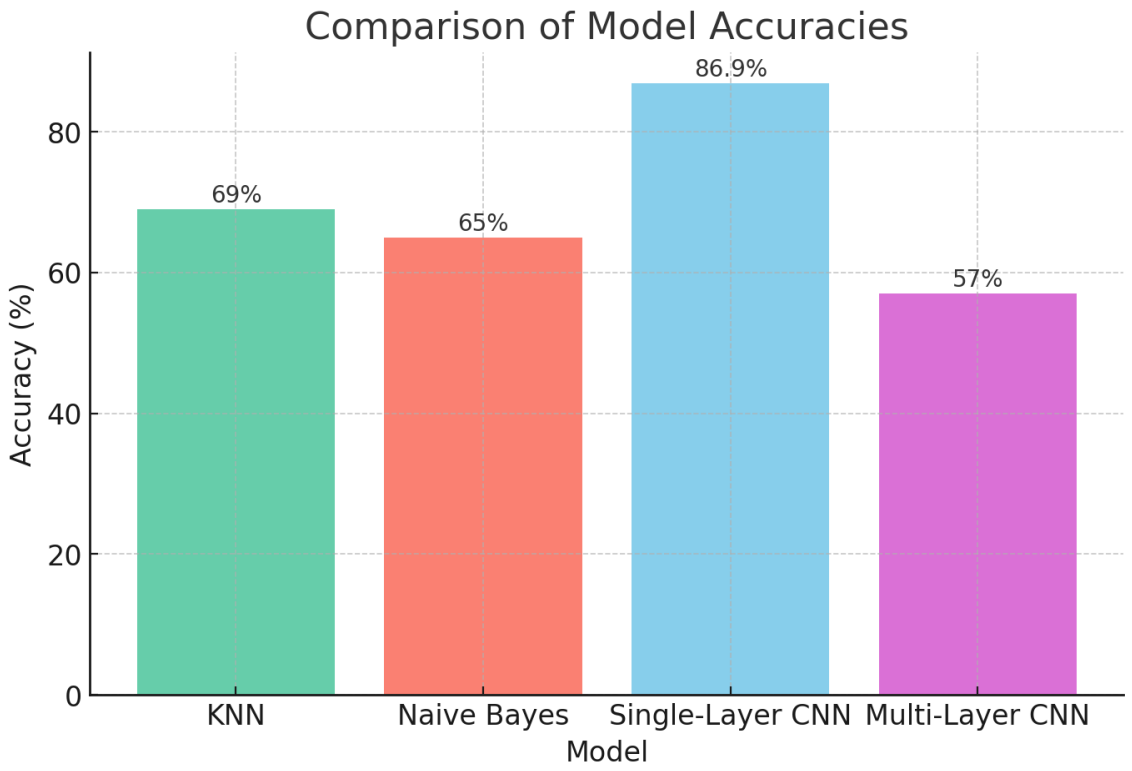
We use a smoothing parameter to avoid zero-probability issues, improve generalization, and prevent overfitting.

Step 3: Utilizing Deep Learning Models (CNNs)

- CNN is particularly suited for extracting spatial features from images, even if they are distorted, translated, or rotated.
- We use two different CNN architectures:
  - Single-Layer: This CNN uses filters and kernels (convolutional layers) to pick up relevant features, which are then passed to a fully-connected layer.
    - We use ReLU for the activation function
    - We also introduce dropout regularization.
  - Multilayer: This architecture increases the number of convolutional layers to capture more complex features.
    - Max-pooling layers are used to downsample and reduce data dimensionality, reducing computational requirements while maintaining important features.
    - Batch normalization normalizes inputs to each layer to improve convergence.

Results

- Of the models, the single-layer CNN performed the best.
  - The multilayer CNN suffered from over-fitting, while KNN and Naïve Bayes had difficulty capturing complex patterns.



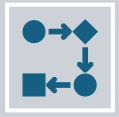
Conclusion

In this study, we proposed a machine-learning pipeline for prostate cancer detection using both classical and deep learning ML approaches.

The data set consisted of low- and high-dimension data.

Classical algorithms are better suited to low-dimensional data while deep learning models perform better on image data.

Provided that the deep learning models balance complexity and accuracy.



In the future, we plan to integrate denoising techniques and subtyping models to improve predictive power and sub-disease group discovery.

Acknowledgements

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.

CALIFORNIA STATE UNIVERSITY
SACRAMENTO