1) Objective. Using multiple machine learning (ML) approaches, we will investigate the effectiveness of computer-aided diagnostic tools in the classification of prostate gland hyperplasia. A binary classification between malignant and benign prostate growth will be used as a proof-of-concept for future classification studies. The aim of the proposed research program is to employ ML tools in the analysis of processed image data from diseased prostate tissue to discriminate between prostate cancer and benign prostatic hyperplasia (BPH). Current paradigms for diagnosing prostate cancer rely on blood tests, MRI, and biopsies. These methods can be time and resource intensive, which is a concern when disease prognosis is heavily dependent on time of detection. A binary classification tool using only image data would lessen the strain placed on healthcare systems by prostate disease and enable quicker diagnoses. This analysis will employ three ML approaches: logistic regression, K-Nearest Neighbours (KNN), and support vector machines (SVM) with both polynomial and radial basis function kernels to classify prostate disease. We expect that logistic regression will outperform the other two approaches because of advantages in modelling binary dependent variables – however, we stress that future applications beyond binary classification may benefit more from KNN or SVM methods.

2) Background. The global burden of prostate cancer is immense and growing [1]. It is the third most diagnosed cancer after lung and breast cancer, and the most common cancer among men [1]. In the United States, nearly 200,000 men are diagnosed each year with over 30,000 deaths [2]. Worldwide, over 1.2 million cases are diagnosed annually[3]. Incidence tends to be most prevalent in the developed world, but rates of incidence have been increasing in the developing world as well [4]. Prognoses vary greatly depending on several factors – chief among them being stage at detection [5]. If detected early enough, 5-year survival rates in the developed world can exceed 98% [6]. If the cancer progresses to later stages or metastasizes then survival rates drop off precipitously to below 33% [6]. Therefore, healthcare strategies placing emphasis on early detection and rapid diagnostics will have the most impact on mortality rate.

There are a few key properties of prostate disease that influence diagnostics and prognosis greatly. Prostate cancer is usually slow growing, and symptoms may not present in the patient until the disease is well established in some cases [7]. There are lymph nodes in proximity of the prostate gland through which the cancer can metastasize. If malignant cells are in close vicinity to draining lymph nodes, the risk of metastasis increases relative to cancerous cells on lymph-distal portions of the prostate gland [8]. The symptoms of prostate cancer include difficulty urinating, the presence of blood in urine, and regional discomfort or pain. These symptoms share high similarity with another prostate disease – benign prostatic hyperplasia (BPH). As implied by its name, BPH is not malignant and non-fatal. It clinically presents as enlargement of the prostate gland and can be detected manually during routine digital rectal examination [9]. BPH increases serum levels of prostate-specific antigen (PSA) which can be detected with conventional blood testing [10]. For this reason, PSA levels alone are not considered discriminatory when diagnosing prostate cancer since they are elevated in both diseases. Instead, worrisome cases will be referred for biopsies, where a final diagnosis can be made. Medical imaging is often done in the pre-operatory period of a biopsy to illuminate areas where tumours are present. MRI's and ultrasounds are the most used forms of imaging to locate prostate cancers, with MRI's offering far superior resolution [11]. Certain regions of the prostate gland develop more easily identifiable malignancies than others. Due to this, lesions cannot easily be distinguished from surrounding tissue is some regions without the use of contrast enhancement and diffusion weighted MRI's [11]. Nonetheless, medical imaging techniques form the basis of tumour mapping done before and during biopsies and therefore serve as a critical pinch-point in prostate cancer diagnostics.

Biopsies are the most critical tool for diagnosing prostate cancers. They are the only deterministic tool available to clinicians for diagnosis, and their purpose goes beyond the simple binary classification of tumour growth as benign or malignant. They are also used to determine the type of cancer the patient has, its grade, stage, and severity [12]. Since many prostate cancers are slow-growing and even non life-threatening, the information gleaned from biopsy is critical for clinicians to make decisions regarding effective treatment strategies. Mistakes made here could prove fatal if appropriately aggressive treatment strategies are not pursued with misdiagnosed high-risk cancers. Despite this, there are worrisome shortcomings in the current standards for biopsies. First, differentiation between benign growth and malignancies have a considerable degree of uncertainty when comparing imaging-guided biopsy methods. For example, clinically significant cancers were found at a higher rate with MRI-guided biopsy than standard ultrasound biopsy [13]. Since MRI machines and the associated tools are expensive and in high demand, the standard of care available to most patients may not give them the most accurate diagnoses. Ultrasound guided biopsies will take multiple cores from different regions of the prostate gland based on intra-operative ultrasound imaging that is performed trans-rectally [12]. Unfortunately, this method is done in a nearly blind fashion as only the larger structure of the prostate gland is visible during the biopsy and low-soft tissue resolution means that malignancies are indistinguishable from surrounding tissues [14]. This is the reason why ultrasound guided biopsies rely on sampling from multiple regions of the prostate gland – it reduces the risk of missing cancerous regions. By comparison, pre-operative MRI imaging both as a stand-alone method of mapping and when used in conjunction with ultrasound produces far better results. Diagnoses made from MRI-guided prostate biopsies have been shown to identify true-positives with 59% accuracy and true-negatives with 84% accuracy [15].Side-by-side comparisons have shown that MRI-guided biopsy improved detection by nearly 20% compared to traditional ultrasound imaging and successfully decreased misdiagnosis of low-risk disease [16].

Despite the high rate of success from MRI-guided biopsy, there are still clearly issues with diagnosis. A specificity of 59%, while superior to ultrasound-guided methods, is below an ideal standard of certainty for such a potentially life-threatening disease. Furthermore, biopsies are invasive procedures and doing them multiple times to confirm that a diagnosis is correct would be both an inefficient use of a clinician's time, and negatively impact the patient's quality of life. Here, machine learning methods applied to MRI data could conceivably ascertain the presence of a malignancy with higher accuracy than a human actor. A high-fidelity machine classifier could augment the findings of a clinician to improve diagnostic accuracy greatly.

3) Methods. Dataset: The dataset used to train all the machine learning models described below was obtained from Kaggle. It included 100 observations of MRI data from diseased prostate glands containing eight clinical features, an ID, and a binary feature indicating whether the final diagnosis for the patient was 'M' (malignant) or 'B' (benign). With a total of just 800 observations of clinical features, the data set was relatively small. Of the 100 observations made, 68 were malignant and 32 were benign. This data set was chosen for its relatively small feature-space and number of observations. For practical purposes, a smaller dataset was preferable for analysis on a personal computer and could be used as a proof-of-concept for scaled up analyses in the future.

Software: All the machine learning methods applied during this preliminary analysis utilized the scikit-learn machine learning library for the Python programming language. The library includes algorithms for k-means clustering, regression and support vector machines. The library can integrate with many other data science Python libraries including NumPy, pandas and matplotlib so that representing outputs from the algorithms with figures is simplified.

Principle Components Analysis (PCA): Though there was a relatively low feature space in the obtained dataset, a PCA was still performed on the data using the matplotlib library. PCA is a statistical tool that can be applied to highly dimensional datasets to inform the user which dimensions account for the greatest sources of variation in the dataset. PCA can thus be used to inform decisions related to dimensionality reduction in machine learning applications. It achieves this by creating artificial variables that are linear combinations of the original variables in the dataset. These linear combinations are the principle components of the dataset, of which there is an upper limit corresponding either to the number of features, or the number of observations – whichever is smaller. Subsequent PC's are always orthogonal to the previous PC. Put briefly, PCA takes some dimensional data and centers it around the origin of its intersecting axes and then finds the line of best fit for the data that passes through the origin, calling this PC1. Then, a line of best fit running orthogonal to the first PC and running through the origin. Subsequent PC's are found in the same manner, each being orthogonal to the last, until the number of PC's matches the number of measured features in the data. Each PC is a linear combination of all the features in the data. Once all the PCs have been discovered, the eigenvalues (the sum of squared distances to the line of best fit) are computed to the determine the proportion of the total variation that each PC accounts for in the data. This information is then used to select features during the analysis of the data with machine learning approaches so that non-relevant dimensions can be removed. In the dataset used for this analysis, PCA was performed and a scree plot visualizing the relative contribution of each PC was provided. This information was subsequently used to reduce dimensionality during analysis with our k-means clustering method.

<u>Logistic Regression:</u> Linear regression methods involve some sort of linearization between correlated data to create a function that can use inputs to interpolate (predict) an output of some continuous data type. In machine learning, this property of linear regression is often used in the form of multiple regression, where multiple parameters are interpreted to predict another property. Logistic regression is like linear regression, except that the former uses input data to predict something with discreet categories like 'True' or 'False'. Rather than fitting a line to the data using the sum of least squares, logistic regression fits an 'S' shaped logistic function to the data using maximum likelihood estimation. Our dataset uses eight clinical features to determine a binary output of ether 'malignant' or 'benign' so using logistic regression is useful because of this advantage in modelling binary dependent variables.

<u>K-Nearest Neighbours:</u> The KNN algorithm first takes some data organized into known categories and then cluster that data using a clustering method of the user's choice. In our case, the training data was a subset of the MRI data obtained from Kaggle and PCA was employed as the clustering method. Then, the testing data is added to the plot and is classified based on their distance to the nearest neighbouring observations already in the PCA plot. Since our data had a binary classification of either malignant or benign, there were two categories of observations to which new data points could have their distance measured. 'K' refers to the number of nearest neighbours that are used to decide on a classification for a new observation. 'K' can be set by the user to provide the most accurate classification. One drawback of this method is that datasets with a high number of categories suffer from poor classification if the number of observations is insufficient to determine a category. This is because the more 'groups' or 'clusters' that exist for a given number of observations, the fewer the number of observations in each cluster. Since final classification is based on which category has the highest number of neighbours to the unknown observation (and not the relative distance to the points) this can skew results at higher values of K. In our case, KNN is a sound approach because there are only two categories and 100 observations.

<u>Support Vector Machine (SVM):</u> We used a support vector machine with both polynomial and radial basis function (RBF) kernel methods to compare ability to classify malignancies correctly. A support vector machine uses a soft margin classifier (also known as a support vector classifier) to set a threshold between two categories to which new observations can be compared for categorization. The shape of the support vector classifier (SVC) depends on the number of dimensions (features) that are included in the categorization. Finding SVC's for data with no clear binary clustering is difficult so kernel functions are used to systematically find SVC's in higher dimensions. The polynomial kernel has a parameter 'd' that stands for the degree of the polynomial. As the value of d increases, the dimension in which the corresponding SVC is found also increases. The RBF kernel finds SVC's in an infinite number of dimensions with behaviour resembling a weighted nearest neighbour model during classification of a new observation where the closest observations in the raining data are highly influential in determining the classification of the new observation.

4) Preliminary Analysis. <u>PCA:</u> PCA plots were generated using the sklearn.decompositon library in Python. A scree plot for the results of the analysis is shown in figure 1. More than 90% of cumulative variance in the data set was explained by the first five principal components, but each of the first six components accounted for significant increases in share of cumulative variance. Examining the scatter and density plots that were provided from Kaggle with the dataset (Figure 2), we can see why certain related features may not have impacted cumulative variance. There were no obvious patterns in the scatter plots to suggest a bimodal distribution of observations to correlate with the clinical diagnosis of malignant or benign. Patterns only existed between the area and perimeter features, which is expected. Results of the PCA were used to inform dimensionality reduction during analysis with KNN.

Logistic Regression: The cost over iteration of the logistic regression model is shown in figure 3. Compared to the other machine learning methods described here, logistic regression had the highest correct classification rate at 87% for the training data and 93% for the testing data. Dimensions were not reduced prior to running the analysis. Sk-learn's implementation of the logistic regression function considers regularization, which prevents overfitting of the data and was not done with the other ML methods presented here. This may have also influenced the success seen with this technique compared with others. Since logistic regression is expressly designed to handle predictions of a categorical variable based on a set of independent variables, some advantages were expected here. To compute this result, the model was trained on 85 of the 100 observations and tested on 15 observations. The low size of the testing sample may have impacted the results, so further modelling with larger datasets is required to confirm these results. Nonetheless, a correct classification rate of 93% contrasts nicely with the specificity of 59% quoted for diagnoses resulting from MRI-guided biopsies of prostate tissue. As a proof-of-concept, logistic regression is a promising candidate for further studies of prostate disease.

KNN: The results of analysis with KNN had an accuracy of 90% on the training data set and 87% on the testing data set. Dimensionality was reduced according to our PCA. Running the KNN algorithm across all PCs showed that the most accurate classifications were obtained when the first three principal components were included in the analysis (Figure 4, panel A). Furthermore, the misclassification error was lowest when the nearest 23 neighbours were used to classify new data points (Figure 5). Though this is a relatively high value for K only two categories exist for the predicted variable, so there was some room error. The model was trained on 67% of the dataset and tested on the remaining 33%. Since the number of observations was very small, this means that there were likely fewer than 23 observations falling into the 'Benign' category during analysis. This is a likely contributing factor to the high number of false positives that were detected with this model (see 'recall' column in figure 4). As a first-run analysis the algorithm performed relatively well despite the false-positive rate. With future improvements, we can expect that KNN will be able to classify prostate disease with higher accuracy that clinicians using MRI data alone.

SVM: Results from the preliminary analysis using SVM had similar success rates with both the polynomial and RBF kernels. The 'C' parameter was varied between successive runs but had little effect on the outcome. 'C' sets the margin size for the boundary between the two categories being predicted but increases beyond C=2 did not improve the accuracy of the model using the RBF kernel and had no effect when using the polynomial kernel. The SVM model utilizing the RBF kernel had a maximum success rate with testing data of 78.8% compared with 75.6% with the polynomial kernel (figure 6 & 7). Low sample size was also a possible contributing factor here as dimensions were not reduced and the testing sample was comprised of just 33 observations. Thus, there were likely few – if any – data points falling within the margins of error set by C resulting in minimal improvements to accuracy by varying this parameter. Many tests of the SVM using the polynomial kernel failed outright or had an unacceptably long runtime. This was likely because the dimensionality of the feature space of a polynomial kernel is quadratic in the number of dimensions in the original input space. Since the input space already had eight dimensions, the runtime increased exponentially at polynomial degrees at three or above. Future analyses done with high performance computing clusters may be able to provide more insight into the usefulness of SVMs with prostate disease data.

5) Conclusion. Despite a low sample size and feature space, all three machine learning models examined in our preliminary analysis had promising rates of correct classification compared to clinical standards with prostate disease. As a proof-of-concept, using machine learning methods could viably improve diagnostics for prostate cancer. At the lowest end, SVM's classified malignant growth correctly in 75.6% of tested samples, exceeding correct diagnosis rates at first biopsy. Moving forward, we could take classification with machine learning beyond a simply binary of malignant vs. benign and use modelling to assist with classifying the grade, stage and type of cancer a patient has using a combination of MRI and biopsy data. In the immediate future, more testing needs to be done to validate preliminary results with datasets including a high number of observations.

6) Figures.

Figure 1.

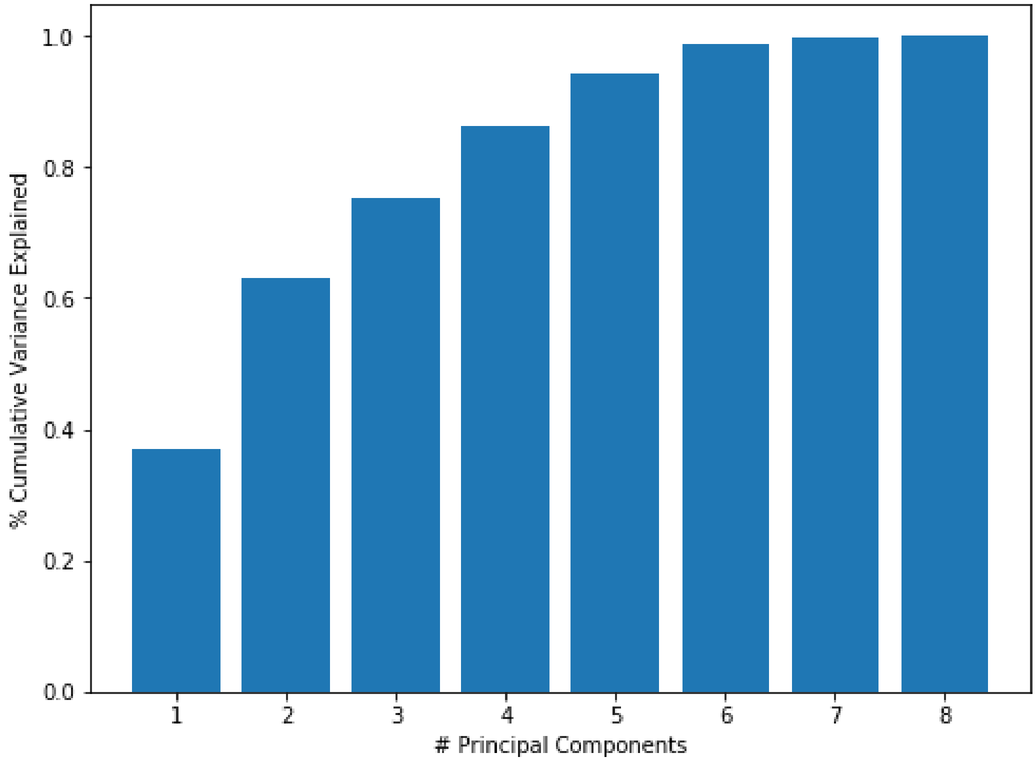


Figure 1 Scree plot of principal component analysis. The first six PC's accounted for nearly 100% of the variation in the dataset.

Figure 2.

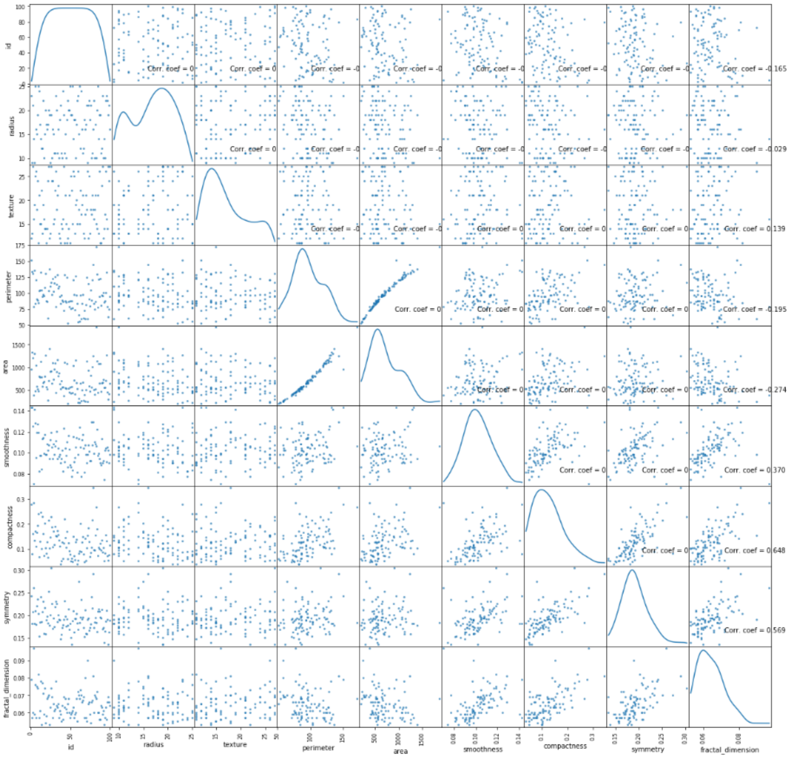
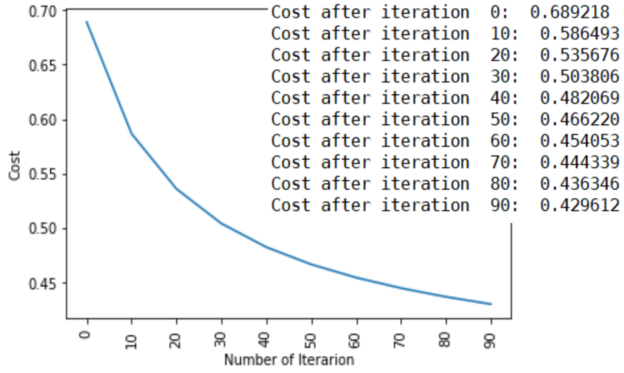
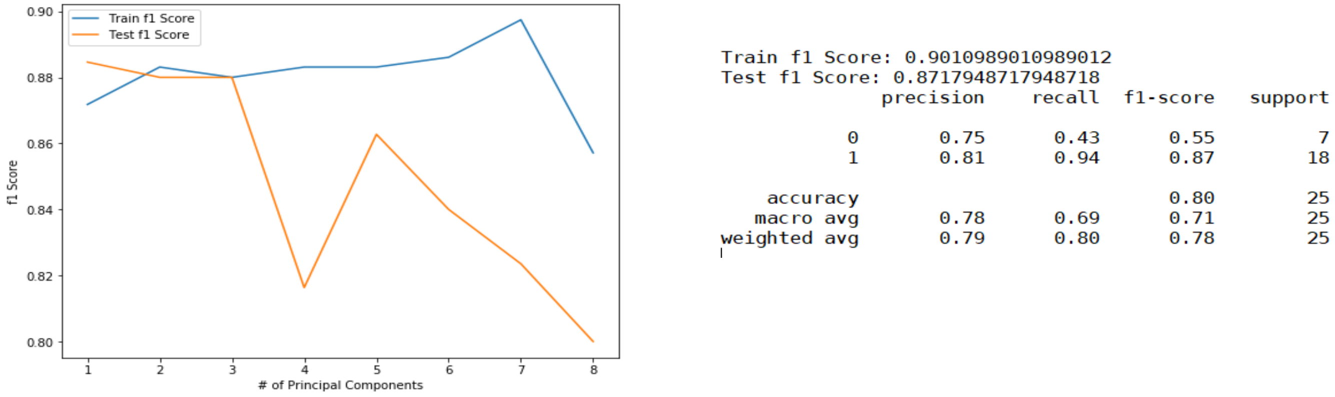


Figure 3.



```
Cost after iteration  0:  0.689218
Cost after iteration  10:  0.586493
Cost after iteration  20:  0.535676
Cost after iteration  30:  0.503806
Cost after iteration  40:  0.482069
Cost after iteration  50:  0.466220
Cost after iteration  60:  0.454053
Cost after iteration  70:  0.444339
Cost after iteration  80:  0.436346
Cost after iteration  90:  0.429612
```

Figure 4.



```
Train f1 Score: 0.9010989010989012
Test f1 Score: 0.8717948717948718
              precision    recall  f1-score   support

           0       0.75      0.43      0.55         7
           1       0.81      0.94      0.87        18

    accuracy                           0.80        25
   macro avg       0.78      0.69      0.71        25
weighted avg       0.79      0.80      0.78        25
```

Figure 5.

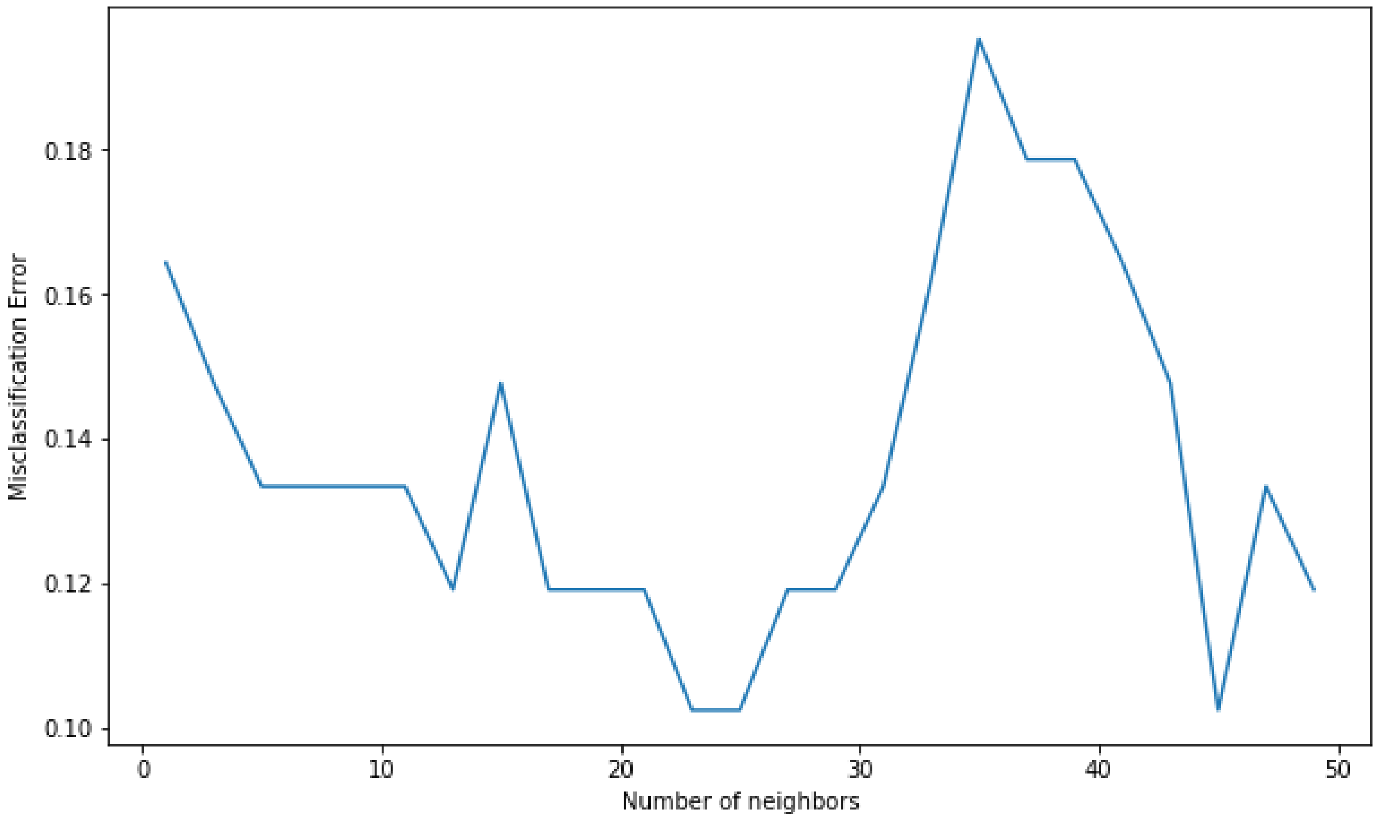
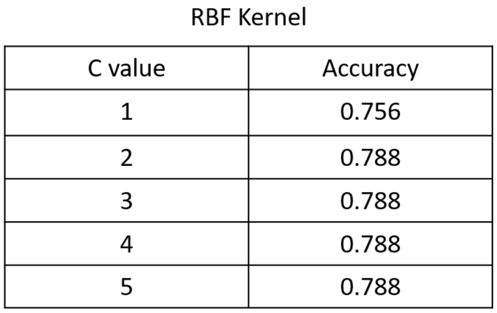


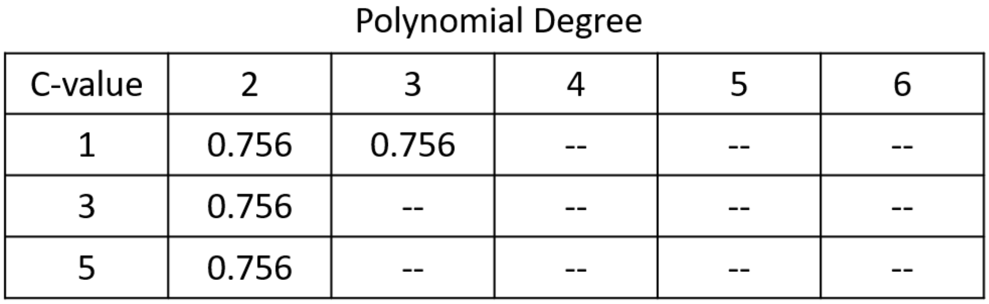
Figure 6.



RBF Kernel

C value	Accuracy
1	0.756
2	0.788
3	0.788
4	0.788
5	0.788

Figure 7.



Polynomial Degree

C-value	2	3	4	5	6
1	0.756	0.756	--	--	--
3	0.756	--	--	--	--
5	0.756	--	--	--	--

6) REFERENCES.

Dataset: https://www.kaggle.com/sajidsaifi/prostate-cancer

1. Baade, P.D., Youlden, D.R. and Krnjacki, L.J. (2009), International epidemiology of prostate cancer: Geographical distribution and secular trends. Mol. Nutr. Food Res., 53: 171-184. https://doi.org/10.1002/mnfr.200700511
2. American Cancer Society: Cancer Facts and Figures 2020. Atlanta, Ga: American Cancer Society, 2020.
3. *Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (November 2018).* "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". *CA: A Cancer Journal for Clinicians.* ***68*** *(6): 394–424. doi:10.3322/caac.21492*
4. Countries With The Highest Incidence Of Prostate Cancer In The World. *World Atlas.* 2020 [Web Article]. Retrieved from: https://www.worldatlas.com/articles/countries-with-the-highest-prevalence-of-prostate-cancer-in-the-world.html
5. *"Prostate Cancer Treatment (PDQ) – Patient Version". National Cancer Institute.* 2014.
6. *"SEER Stat Fact Sheets: Prostate Cancer".* SEER 2020. Retrieved from: https://seer.cancer.gov/statfacts/html/prost.html
7. *"Prostate Cancer Treatment (PDQ) – Health Professional Version". National Cancer Institute.* 2014 [PDF]
8. Swanson GP, Hubbard JK. A better understanding of lymphatic drainage of the prostate with modern imaging and surgical techniques. Clin Genitourin Cancer. 2013 Dec;11(4):431-40. doi: 10.1016/j.clgc.2013.04.031. Epub 2013 Jun 29. PMID: 23820065.
9. Kim, EH; Larson, JA; Andriole, GL (2016). *"Management of Benign Prostatic Hyperplasia". Annual Review of Medicine (Review).* ***67****: 137–51.*
10. Chang, RT; Kirby, R; Challacombe, BJ (April 2012). *"Is there a link between BPH and prostate cancer?". Practitioner.* ***256*** *(1750): 13–6, 2.*
11. Georgiev, A. (2016). Case of prostate cancer with anterior localization - Multiparametric MRI study. Rentgenologiya i Radiologiya, 55(4), 285–287.
12. Ghei M, Pericleous S, Kumar A, Miller R, Nathan S, Maraj BH (September 2005). "Finger-guided transrectal biopsy of the prostate: a modified, safer technique". *Annals of the Royal College of Surgeons of England.* ***87*** *(5): 386–7.*
13. Kasivisvanathan V, Rannikko AS, Borghi M, Panebianco V, Mynderse LA, Vaarala MH, et al. (May 2018). "MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis". *The New England Journal of Medicine.* ***378*** *(19): 1767–1777. doi:10.1056/NEJMoa1801993*
14. Bonekamp D, Jacobs MA, El-Khouli R, Stoianovici D, Macura KJ (May–June 2011). "Advancements in MR imaging *of the prostate: from diagnosis to interventions". Radiographics.* ***31*** *(3): 677–703.* doi:10.1148/rg.313105139
15. Isebaert S, Van den Bergh L, Haustermans K, Joniau S, Lerut E, De Wever L, et al. (June 2013). "Multiparametric MRI for prostate cancer localization in correlation to whole-mount histopathology". *Journal of Magnetic Resonance Imaging*. **37** (6): 1392–401. *doi*:*10.1002/jmri.23938*
16. Pokorny MR, de Rooij M, Duncan E, Schröder FH, Parkinson R, Barentsz JO, Thompson LC (July 2014). "Prospective study of diagnostic accuracy comparing prostate cancer detection by transrectal ultrasound-guided biopsy versus magnetic resonance (MR) imaging with subsequent MR-guided biopsy in men without previous prostate biopsies". *European Urology*. **66** (1): 22–9. *doi*:*10.1016/j.eururo.2014.03.002*