PLOS ONE

RESEARCH ARTICLE

Discriminatory Gleason grade group signatures of prostate cancer: An application of machine learning methods

Mpho Mokoatle[1☯]***, Darlington Mapiye**[2☯]**, Vukosi Marivate**[1,3☯]**, Vanessa M. Hayes**[3,4‡]**, Riana Bornman**[4‡]

1 Department of Computer Science, University of Pretoria, Pretoria, South Africa, **2** AstraZeneca, London, United Kingdom, **3** School of Medical Sciences, The University of Sydney, Sydney, Australia, **4** School of Health Systems and Public Health, University of Pretoria, Pretoria, South Africa

☯ These authors contributed equally to this work.
‡ VMH and RB also contributed equally to this work.
* u19394277@tuks.co.za

Abstract

One of the most precise methods to detect prostate cancer is by evaluation of a stained biopsy by a pathologist under a microscope. Regions of the tissue are assessed and graded according to the observed histological pattern. However, this is not only laborious, but also relies on the experience of the pathologist and tends to suffer from the lack of reproducibility of biopsy outcomes across pathologists. As a result, computational approaches are being sought and machine learning has been gaining momentum in the prediction of the Gleason grade group. To date, machine learning literature has addressed this problem by using features from magnetic resonance imaging images, whole slide images, tissue microarrays, gene expression data, and clinical features. However, there is a gap with regards to predicting the Gleason grade group using DNA sequences as the only input source to the machine learning models. In this work, using whole genome sequence data from South African prostate cancer patients, an application of machine learning and biological experiments were combined to understand the challenges that are associated with the prediction of the Gleason grade group. A series of machine learning binary classifiers (XGBoost, LSTM, GRU, LR, RF) were created only relying on DNA sequences input features. All the models were not able to adequately discriminate between the DNA sequences of the studied Gleason grade groups (Gleason grade group 1 and 5). However, the models were further evaluated in the prediction of tumor DNA sequences from matched-normal DNA sequences, given DNA sequences as the only input source. In this new problem, the models performed acceptably better than before with the XGBoost model achieving the highest accuracy of 74 ± 01, F1 score of 79 ± 01, recall of 99 ± 0.0, and precision of 66 ± 0.1.

OPEN ACCESS

Citation: Mokoatle M, Mapiye D, Marivate V, Hayes VM, Bornman R (2022) Discriminatory Gleason grade group signatures of prostate cancer: An application of machine learning methods. PLoS ONE 17(6): e0267714. https://doi.org/10.1371/journal.pone.0267714

Editor: Sathishkumar V E, Hanyang University, KOREA, REPUBLIC OF

Received: December 16, 2021

Accepted: April 13, 2022

Published: June 9, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0267714

Copyright: © 2022 Mokoatle et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data cannot be shared publicly because of ethical reasons since it is human samples. Data are available from the University of Pretoria (contact via u19394277@tuks.co.za) for researchers who meet the criteria for access to confidential data.

Funding: The work reported herein was made possible through funding by the South African Medical Research Council (SAMRC) through its Division of Research Capacity Development under the Internship Scholarship Program from funding received from the South African National Treasury. The content hereof is the sole responsibility of the authors and does not necessarily represent the official views of the SAMRC or the funders. The award recipient is Mpho Mokoatle. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

Prostate cancer is the leading male cancer in South Africa and is the second most frequently diagnosed cancer among men globally [1]. As men live longer, there is an increase in the occurrence and mortality of the disease [2]. Except for age, the main risk factor is hereditary. Other factors such as race, high-calorie diet, and exposure to heavy metals have a significant impact on the risk of occurring the disease [3, 4].

When it comes to the diagnosis of prostate cancer, a prostate biopsy procedure is common [5]. This procedure involves the extraction of tissue samples from the prostate by using specialised biopsy needles. It is typically performed by using an ultrasound probe that is placed in the rectum which than produces a real-time image of the prostate. The samples produced from this procedure are then taken to a pathologist for evaluation and grading [6, 7].

The Gleason grade group system is the most reliable method and criterion for selection of therapy. In 2014, the International Society of Urological Pathology (ISUP) [8] released supplementary guidance on an improved prostate cancer grading system called the ISUP-Grade Group. This system is simpler, with just five grades, 1 to 5, to describe the growth of the tumor. Grade 1 refers to the least aggressive growth of the tumor, and grade 5 refers to the most aggressive growth [9].

Due to the difficulty and natural subjectivity of this system, Gleason grading is affected by large discordance rates among pathologists (30-50%) [10–15]. However, grades provided by experts with numerous years of experience are more accurate and precise more than grades provided by pathologists with only a few years of experience [16–19], indicating the need to improve the clinical usefulness of the system by improving grading discordance and accuracy [20].

In this work, the DNA sequences that were sequenced from patients that present with a Gleason grade 1 and 5 are studied. The objective of this work is to find discriminatory features within the DNA sequences, and map them to their correct Gleason grade group using machine learning. Two key cancer genes are investigated: *BRCA 1* and *BRCA 2*. These genes have been key genes of interest in prostate cancer [21]. Studies that interrogated these two genes suggest that men who harbor a disease-associated BRCA 2 allele have an increased predisposition of prostate cancer (2 to 5-fold increased risk). This finding suggests that deleterious mutations in BRCA 2 play a significant role in the susceptibility of prostate cancer [22, 23]. Different from BRCA 2 mutations, mutations in BRCA 1 have been inconsistently correlated with the risk of prostate cancer. Studies that have evaluated prostate cancer risk in men that carry BRCA 1 mutations have reportedly been negligible, but not insignificant [24, 25]. The contributions of this work are summarised as follows:

- this study specifically compares two extremes of the Gleason grade group (Gleason grade group 1 and 5)
- while previous studies have used medical images and clinical features [26–31] as input to their Gleason grade group predictor models, this study explores the challenges that are encountered when blood DNA sequences are used as the only input source to the machine learning models.

This work is divided as follows: first, a literature review will be given that highlights the gap in the prediction of the Gleason grade group in the context of machine learning. Second, the data and description of methods will be discussed. Finally, the results, discussion, and conclusion section will follow.

2 Literature survey

Recently, deep learning has emerged as a powerful tool to automate the Gleason Grading system. Deep learning systems make use of multi-faceted neural networks that are able to extract complex features from data. Recent work [26] designed a Gleason score annotator by using a convolutional neural network (MobileNet) on tissue microarrays images. The final output layer of this architecture produced a probability distribution over four possible Gleason classes. A key limitation in this work is that the training, testing, and validation sets were too small, which led to some bias in the predictions produced by the model. A recent study [27] similar to this one also used a convolutional neural network (Inception V3) to develop a Gleason score annotator using whole slide images. In addition to predicting a Gleason pattern, this architecture first provided a probability distribution over an image being benign or malignant. Different from the above work, a study [28] applied a convolutional neural network on multiparametric magnetic resonance imaging (mpMRI) images of prostate cancer patients to extract deep entropy features. Then, the features extracted from the convolutional neural network were used as input to a Random Forest model for prediction of the Gleason grade group. Even though the training data was too small, the performance measure would have been more reliable if the models were cross-validated. Biopsy images of patients who underwent a prostate biopsy following suspicion of prostate cancer has also been used as input to convolutional neural networks (U-Net and an Inception-v3 Network) for the prediction of the Gleason grade group and cancer detection [29, 30]. To validate the performance of the deep learning system, the predictions from the models were compared with those of pathologists where a high agreement was found between the deep learning systems and the pathologists.

Unlike using convolutional neural networks for the prediction of the Gleason grade group, a study [31] developed a machine learning assisted model that predicts the probability of a patient having a Gleason grade upgrade before treatment. The input used to the machine learning models (Logistic Regression, Random Forest, Support Vector Machine) were clinical features such as age, prostate-specific antigen (PSA) level, and the clinical stage.

Overall, much emphasis has been placed on creating machine learning models that predict the Gleason grade group from medical images and clinical data. To the best of our knowledge, this is the first study that focuses on DNA sequences as the only input source to a Gleason grade group prediction model. This work explores the challenges that are associated with finding discriminatory signatures within the DNA sequences of patients that present with a Gleason grade group of 1 and 5.

3 Data description, data representation methods, machine learning algorithms, and sequence similarity

3.1 Data description

Patients were recruited and consented according to approval granted from the University of Pretoria Faculty of Health Sciences Research Ethics Committee 43/2010 (South Africa); DNA sequencing was generated under approval granted from the St. Vincent's Hospital Human Research Ethics Committee (HREC) SVH/15/227 in Sydney (Australia), and this study was approved by the Faculty of Engineering, Built Environment & IT (Ethics Reference No: 43/2010; 11 August 2020). The data was fully anonymized before analysis.

The DNA sequences of twelve patients with a histopathological ISUP-GG of 1 (low risk prostate cancer) and 5 (high-risk prostate cancer) were selected for analysis. The DNA sequences were aligned using the BWA-MEM aligner [32] to produce BAM files. The BAM files were converted to FASTA files using samtools [33] and an in-house python script was used for pre-processing and removing IDs from the blood DNA sequences.

The blood DNA sequences were then truncated into *k*-mers. *k*-mers are defined as all the possible substrings of length *k* that are contained in a sequence [34]. The classification problem in this work is defined as follows: given a DNA sequence *x* that consists of *k*-mers of size 63, can a machine learning function *f* learn the correct mapping from the input *x* to the outcome variable *y* (Gleason grade group of 5 or 1):

$$y = f(x) \tag{1}$$

After preprocessing, the data was transformed into the below data structure (Fig 1).

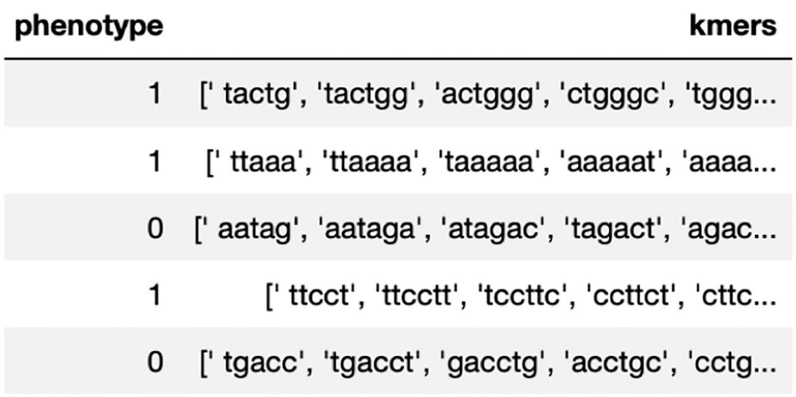
3.2 Data representation methods

To vectorize the *k*-mers, the Term Frequency—Inverse Document Frequency (TF-IDF) [35] algorithm were used. TF-IDF is a statistical method that calculates how significant a token or word is to a document in a set of documents. Two matrices are used to calculate the TF-IDF score: term frequency (TF), which is a measure of how many times a token appears in a document and inverse document frequency (IDF), is a measurement of how frequent or rare a token is in the entire document set. Multiplying these two measurements produces a TF-IDF score of each word in the document [36]. The main disadvantage of TF-IDF is that it produces extremely high dimensional vectors [37]. To overcome this, the Principal Component Analysis (PCA) [38] was used as a data reduction technique to transform the high dimensional vectors into 2-dimensional (d) vectors.

The other vectorization method that was used was the Skip-gram method from the *word2-vec* algorithm. This vectorization method was chosen as it has been found to be robust with regards to transforming DNA or genomic data into dense vector representations in preparation for machine learning [39–42]. In the context of this work, the usefulness of the Skip-gram model lies in determining *k*-mers that are important in predicting the surrounding *k*-mers in a DNA sequence. Precisely, given a sequence of training *k*-mers $w_1, w_2, w_3, \ldots, w_T$ the training objective of the Skip-gram model is to maximise the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logp(w_{t+j}|w_t) \tag{2}$$

where *c* is the size of the context *k*-mers in the training set.



phenotype	kmers
1	[' tactg', 'tactgg', 'actggg', 'ctgggc', 'tggg...
1	[' ttaaa', 'ttaaaa', 'taaaaa', 'aaaaat', 'aaaa...
0	[' aatag', 'aataga', 'atagac', 'tagact', 'agac...
1	[' ttcct', 'ttcctt', 'tccttc', 'ccttct', 'cttc...
0	[' tgacc', 'tgacct', 'gacctg', 'acctgc', 'cctg...

Fig 1. Blood DNA sequences *x* transformed into *k*-mers with their corresponding Gleason grade group *y*.

https://doi.org/10.1371/journal.pone.0267714.g001

In this work, the Skip-gram *k*-mer tokens were represented by a continuous vector of size 100, and summed up with other vectors of the same sequence to give a single continuous vector that represents the entire sequence.

3.3 Machine learning algorithms

After obtaining the 2-d TF-IDF vectors from PCA, they were used as features to several machine learning models: Gradient boosting algorithm: eXtreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Random Forest (RF). XGBoost is an ensemble boosting learning method that makes use of several learners to make predictions. This method is different from other ensemble methods as it builds a sequence of originally weak models into progressively more powerful models, where the errors made by previous models are corrected in subsequent models [43]. The steps involved in the ensemble technique are as follows:

- first, an initial model F_o is initialised to predict the target variable y. This model produces a residual error $(y - F_o)$.
- next, an additive learner h_1 is fit onto the the residuals from the previous step.
- than, F_o and h_1 are summed to produce f_1, which is the boosted version of f_o. The residual error from f_1 will be lower in comparison to the residuals of f_o:

$$F_1(x) \Leftarrow F_o + h_1(x) \tag{3}$$

To improve the performance of f_1, the residuals of f_1 can be modeled to create a new model f_2:

$$F_2(x) \Leftarrow F_1(x) + h_2(x) \tag{4}$$

This procedure can be performed for a few iterations m until residual errors have been minimised as much as possible:

$$F_m(x) \Leftarrow F_{m-1}(x) + h_m(x) \tag{5}$$

Instead of fitting the additive learners $hm(x)$ on the residuals, fitting it on the gradient of the loss function makes this process more generic and applicable across all loss functions. Hence, XGBoost uses the gradient descent algorithm to minimise the loss [43, 44].

The LSTM and GRU are variants of Recurrent Neural Networks (RNNs) that regulate information through the network by using several gates. The gates regulate the flow of information by learning which timestamps are important to keep or discard [45].

In an LSTM cell (Fig 2) the sigmoid function called the forget gate is responsible for deciding which information will be discarded from the cell state. This gate takes as input x_t and the previous hidden state h_{t-1}, than outputs a value between 0 and 1 for each value in the cell state C_{t-1}. If the value is 1, the information from the previous hidden state will be kept and if the value is 0, the information from the previous previous hidden state will be discarded [46–48]:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{6}$$

Next, the input gate i_t has to determine which new information will be added in the cell state. Than, a *tanh* layer will create a vector of new candidate values $\tilde{C}_t$, that will be added to

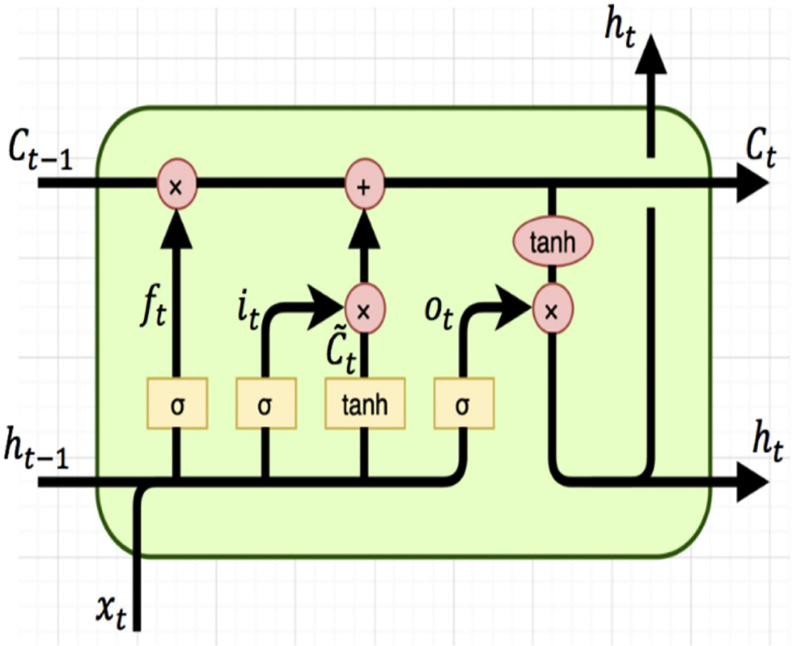


Fig 2. Architecture of an LSTM unit [59].

https://doi.org/10.1371/journal.pone.0267714.g002

the cell state:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{7}$$

$$\tilde{C} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{8}$$

To update the old cell state C_{t-1} into the new cell state C_t, the old state is multiplied by f_t. Next, $i_t * \tilde{C}_t$ is added, which are the new candidate values:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{9}$$

Finally, the output that is based on the cell state is given. The cell state is put through a *tanh* function and multiplied by the output of the sigmoid gate: [46–48]

$$o_t = \sigma(W \circ [h_{t-1}, x_t] + b_\circ) \tag{10}$$

$$h_t = o_t * tanh(C_t) \tag{11}$$

A three layer LSTM architecture was selected with a total of 224 hidden units. The output layer consisted of a sigmoid activation function that provides a probability distribution of a sequence either belonging to a patient with a Gleason grade group of 1 or 5. The training dataset was divided over 50 batches and trained over 5 epochs. Dropout rate at 60% was used to control overfitting of the model.

GRUs (Fig 3) are similar to LSTMs in that they both use gates to regulate the flow of information. GRUs are faster to train than LSTMs, and also have have a simpler architecture [49–51].

Inside a GRU cell, at each timestamp t, the cell takes an input X_t and the hidden state h_{t-1} from the previous timestamp. Next, the cell will output a new hidden state h_t which will be fed

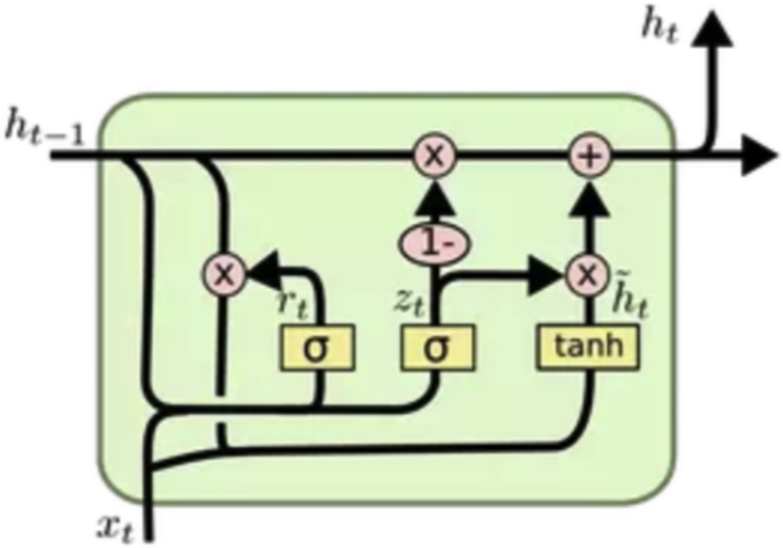


Fig 3. Architecture of a GRU unit [59].

https://doi.org/10.1371/journal.pone.0267714.g003

as input to the next timestamp. Unlike the LSTM that has three gates, the GRU has two gates: the update gate and the reset gate. The reset gate r_t is in charge of the short-term memory of the network. It is responsible for deciding which timestamps to discard [49–51]:

$$r_t = \sigma(X_t * U_r + H_{t-1} * W_r) \tag{12}$$

r_t will output a value between 0 and 1 due to the sigmoid function. As previously mentioned, if the output value is equal to 1, this means that the timestamps from the previous hidden state h_{t-1} will be kept. And if the output value is 0, the timestamps from the previous hidden state h_{t-1} will be discarded [49–51].

To generate the hidden state $\hat{H}_t$ of a GRU cell, a two-step process is followed. First, a candidate hidden state $\hat{H}_t$ needs to be generated:

$$\hat{H}_t = tanh(X_t * U_g + (r_t \circ H_{t-1}) * W_g) \tag{13}$$

the input X_t and the hidden state from the previous timestamp H_{t-1} are multiplied by the output of the reset gate r_t. Next, this is passed to a *tanh* function which outputs the candidates hidden state $\hat{H}_t$. The usefulness of this equation is important in showing how the value of the reset gate is used to control how much influence the previous hidden state can have on the candidate state [49–51].

Similarly, the GRU cell also has an update gate which is responsible for determining how much past information needs to be kept:

$$u_t = \sigma(X_t * U_u + H_{t-1} * W_u) \tag{14}$$

This equation is similar to the one used by the reset gate, the only key differences are the new weight matrices U_u and W_u [49–51].

The GRU models were configured with a stack of four hidden layers and a total of 240 hidden units. The output layer was also a dense layer with a sigmoid activation function, and the model was trained over 5 epochs with the training set divided over 50 batches. Dropout (at

60%) was also used to control overfitting. All the machine learning models were validated via a Repeated *k*-fold cross validation (cv) (cv = 5, runs = 5). The experiments in this work were conducted on a NVIDIA Tesla P100 GPU virtual machine with 100 GB of memory.

RF was also used to find discriminatory signatures between Gleason grade group 1 and 5 blood DNA sequences. In RF, several decision trees are created simultaneously. In the final prediction, the multiple decision trees are merged in order to determine the final answer, which will be the average of all the decision trees [52]. To decide how the nodes of the decision trees would branch, the default Gini index was used:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{15}$$

Where p_i is the relative frequency and *c* represents the number of classes. This equation makes use of the class and probability to determine the Gini of each branch on a node [52].

Another binary machine learning model that was used was the Logistic Regression (LR). The Skip-gram *k*-mer features were used as input to a Logistic Regression (LR) model. A logistic regression model is a machine learning model that uses a decision boundary to separate a set of data points into their distinct classes. A logistic regression is comparable to linear regression, the key difference between them is that logistic regression is used when the target variable is categorical, while linear regression is used when the target variable is continuous. In this study, the target variable is categorical (1 = Gleason grade group of 5, 0 = Gleason grade group of 1). Logistic regression uses a Sigmoid function to convert the probability values *z* to be in the range between 0 and 1:

$$S(z) = \frac{1}{1 + \epsilon^{z}} \tag{16}$$

This function transforms $-\infty$, 0 and $+\infty$ to 0, 0.5, and 1 respectively. If the probability value *z* for a data point is close to $+ \infty$, this is an indication that the data point is above the decision boundary, hence it will belong to the positive class. In contrast, If the the probability value *z* for a data point is close to $-\infty$, it means that the data point is below the decision boundary, meaning it belongs to the negative class. If the data point is predicted to be on the decision boundary, the value of *z* is 0, and the Sigmoid function will transform it to 0.5, meaning that it has a 50% probability of belonging to the positive class [53, 54].

3.4 Sequence similarity

Multicollinearity is a problem in machine learning where two or more predictor variables are highly correlated with each other [55]. This presents a problem because the individual effects of the predictor variable on the target variable would not be distinguishable. One of the methods that is applied to deal with multicollinearity in machine learning is to remove the collinear variables. In the context of this work, removing collinear *k*-mers would result in a completely new set of DNA sequences since the sequences would have to be truncated either in the beginning, middle, or at the end. In the context of this work, multicollinearity can also be equated to sequence similarity in genomics. Sequence similarity is an important concept in genomics that refers to the degree of similarity between sequences [56]. This is often indicated as a percentage of identical bases over a given length of the alignment. The Basic Local Alignment Search Tool (BLAST) was used to evaluate the similarity between blood DNA sequences [57]. When a sequence similarity test is performed between a pair of sequences, several attributes are returned such as the *E value, query cover*, and *percent identity*. In this work, only the *percent identity* is reported. The percent identity refers to how similar the query sequence is to the subject sequence. Specifically, it describes the number of bases that are identical in the sequences. A significant match is 100% [58].

A figure (Fig 4) has been generated to provided an overview of all the methods that were used in this work.

4 Results and discussion

4.1 Sequence similarity results and TF-IDF Visualizations

For both *BRCA 1* and *BRCA 2*, the results (Tables 1 and 2) illustrate that most sequences are highly similar with a percent identity of 90-100%. The lowest percent identity across the sequences is 70-80%, which is still too high. This indicates that blood DNA sequences that are derived from patients that present with Gleason grade group of 5 are not that very different from patients that present with a Gleason grade group of 1. There might exist a small region of dissimilarity, however, at this stage, the number of sequences available for this experiment are inadequate to capture the region of dissimilarity. It is probable that hundreds of thousands of DNA sequences are required to capture this region.

Next, the impact of this high similarity is investigated in the machine learning models to determine if discriminatory signatures (region of dissimilarity) within the DNA sequences can be detected and mapped to their correct Gleason grade group.

To ensure that the machine learning models are trained on distinct sequences, highly similar sequences were removed using BLAST. Before the removal of highly similar sequences, the total number of blood DNA sequences from the *BRCA 1* gene were 235 711. For *BRCA 2*, the total number of the sequences were 243 822. After the removal of highly similar sequences, the table (Table 3) shows the new data distribution and the total number of sequences in each class. Blood DNA sequences that shared more than 25 bases of homology were considered as similar and were thus removed.

In keeping with the high sequence similarity observation amongst the blood DNA sequences as shown above, the TF-IDF visualisation of the *k*-mers (Figs 5 and 6), also show that there is a great overlap between the *k*-mer features of the two Gleason grade groups as no separable clusters were detected.

4.2 Machine learning results

The RF model achieved the highest accuracy as shown (Table 4). However, the recall was too high. This is an indication that the majority of the DNA sequences were predicted as positive

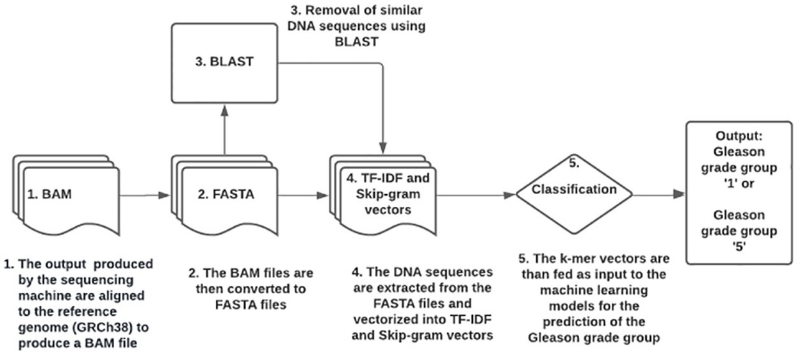

3. Removal of similar DNA sequences using BLAST
3. BLAST
1. BAM
2. FASTA
4. TF-IDF and Skip-gram vectors
5. Classification
Output: Gleason grade group '1' or Gleason grade group '5'
1. The output produced by the sequencing machine are aligned to the reference genome (GRCh38) to produce a BAM file
2. The BAM files are then converted to FASTA files
4. The DNA sequences are extracted from the FASTA files and vectorized into TF-IDF and Skip-gram vectors
5. The k-mer vectors are than fed as input to the machine learning models for the prediction of the Gleason grade group

Fig 4. This figure represents the summary of all the methods that were executed in this work.

https://doi.org/10.1371/journal.pone.0267714.g004

Table 1. Sequence similarity within a Gleason grade group of 5 and 1 for *BRCA 1* blood DNA sequences.

	Grouped by percentage of identical matches	Total no. of local alignments
Gleason grade group 5	90-100	7170891
	80-90	3685304
	70-80	62500
Gleason grade group 1	90-100	7270628
	80-90	3732281
	70-80	56560

https://doi.org/10.1371/journal.pone.0267714.t001

Table 2. Sequence similarity within a Gleason grade group of 5 and 1 for *BRCA 2* blood DNA sequences.

	Grouped by percentage of identical matches	Total no. of local alignments
Gleason grade group 5	90-100	6256450
	80-90	910123
	70-80	17970
Gleason grade group 1	90-100	6510144
	80-90	932427
	70-80	16167

https://doi.org/10.1371/journal.pone.0267714.t002

Table 3. Data count and distribution of classes after the removal of highly similar DNA sequences.

	Gleason grade group 5	Gleason grade group 1
BRCA 1	3111 ~ 58%	2210 ~ 42%
BRCA 2	3108 ~ 62%	1941 ~ 38%

https://doi.org/10.1371/journal.pone.0267714.t003

(Gleason grade group 5), with very few true negatives (Fig 7). This trend was also observed with the other models as well, which is an indication that not enough learning was achieved.

Considering the results of the *BRCA 2* gene (Table 5), the LR and GRU models achieved the highest accuracy while having the highest recalls indicating that a large number of sequences were predicted as positive. The confusion matrix of the GRU model is shown (Fig 8).

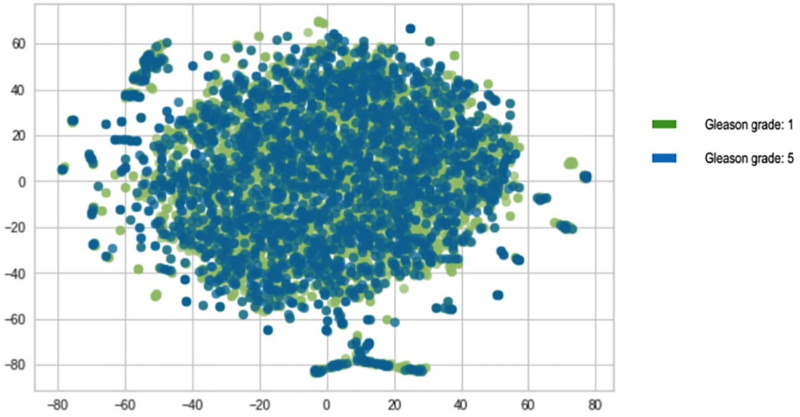


Fig 5. Visualisation of TF-IDF *kmers* for *BRCA 1*.

https://doi.org/10.1371/journal.pone.0267714.g005

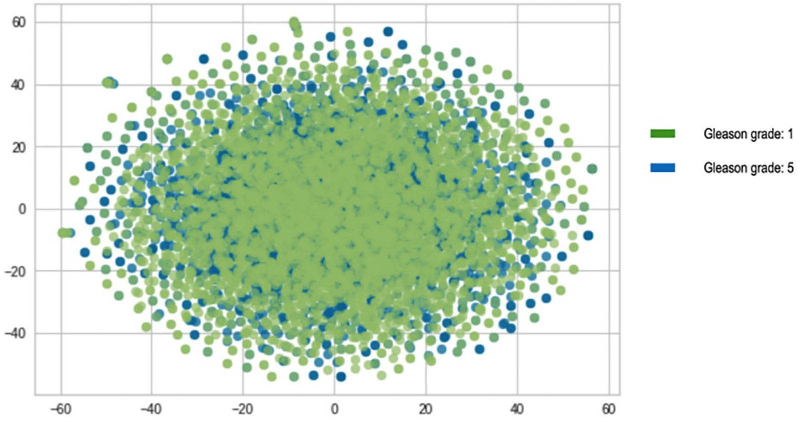


Fig 6. Visualisation of TF-IDF *kmers* for *BRCA 2 kmers*.

https://doi.org/10.1371/journal.pone.0267714.g006

Table 4. This table shows the results of the machine learning models using data from the *BRCA 1* gene.

	Acc (%)	F1 (%)	Recall (%)	Precision (%)
XGBoost	57 ± 1.6	69 ± 1.3	85 ± 2.0	58 ± 1.8
LSTM	58 ± 1.5	74 ± 1.3	100 ± 0.0	58 ± 1.5
GRU	58 ± 1.1	74 ± 0.9	100 ± 0.0	58 ± 1.1
LR	58 ± 1.7	73 ± 1.3	98 ± 0.7	58 ± 1.6
Random Forest	59 ± 1.7	74 ± 1.4	98 ± 0.8	59 ± 1.7

https://doi.org/10.1371/journal.pone.0267714.t004

While some of machine learning models achieved just above average performance, they all seemed to classify most blood DNA sequences as positive (Gleason grade group 5), which suggests that no discriminatory signatures were discovered within the blood DNA sequences of patients that present with a Gleason grade group of 5 and Gleason grade group of 1. This finding further stipulates that are still a lot of opportunities for improvement with regards to designing more robust data representation methods and machine learning classifiers that

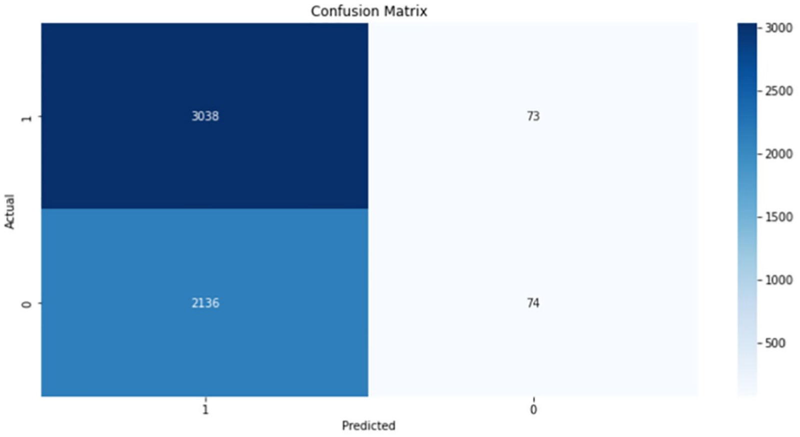


Fig 7. Confusion matrix of the Random Forest model for *BRCA 1*.

https://doi.org/10.1371/journal.pone.0267714.g007

Table 5. This table shows the results of the machine learning models using data from the *BRCA 2* gene.

	Acc (%)	F1 (%)	Recall (%)	Precision (%)
LSTM	58 ± 1.5	73 ± 1.3	100 ± 0	58 ± 1.6
XGBoost	61 ± 1.3	74 ± 1	93 ± 1.3	62 ± 1.4
Random Forest	61 ± 0.1	75 ± 0.8	99 ± 0.6	61 ± 1.1
LR	62 ± 1.3	76 ± 0.1	99 ± 0.2	62 ± 1.3
GRU	62 ± 1.2	77 ± 0.9	100 ± 0	62 ± 1.2

https://doi.org/10.1371/journal.pone.0267714.t005

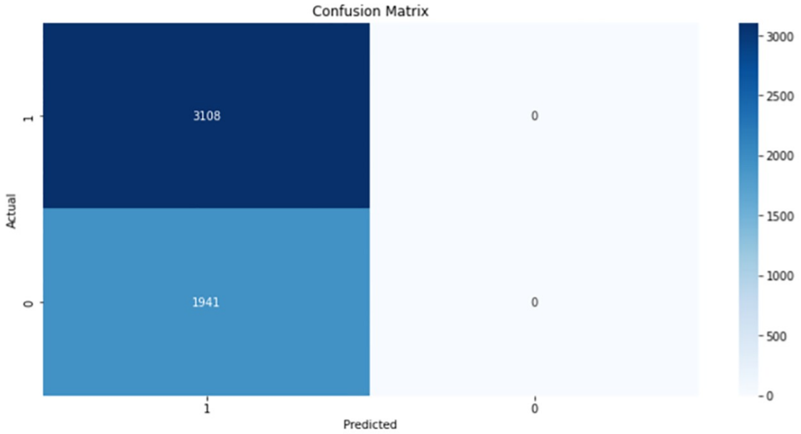


Fig 8. Confusion matrix of the GRU model for *BRCA 2*.

https://doi.org/10.1371/journal.pone.0267714.g008

are adequately sensitive to detect discriminatory Gleason grade groups signatures in DNA sequences.

4.3 Prediction of tumor DNA sequences

Having observed that the above machine learning models were not able to adequately find discriminatory signatures in the DNA sequences of the two Gleason grade groups, a new classification question was formulated: *Given tumor and matched-normal DNA sequences, can the models predict tumor DNA sequences?*. This new problem was formulated to further assess the usefulness of the machine learning models and determine if other classification problems can be learned using DNA sequences as the only input source to the models. In addition, a bigger dataset was used that contained 304 450 tumor DNA sequences and 305 214 matched-normal DNA sequences from the *APC* gene of colorectal cancer patients.

The three machine learning models (LR, RF, and XGBoost) were evaluated to establish if they can distinguish tumor DNA sequences from normal DNA sequences. The results (Table 6) show an overall improvement in the performance of the models compared to the results seen in the previous section of the prediction of the Gleason grade group. In the

Table 6. This table shows the results of the machine learning models using data from the *APC* gene.

	Acc (%)	F1 (%)	Recall (%)	Precision (%)
LR	65 ± 0.1	67 ± 0.1	71 ± 0.1	63 ± 0.1
Random Forest	71 ± 0.1	75 ± 0.3	87 ± 0.3	66 ± 0.3
XGBoost	74 ± 0.1	79 ± 0.1	99 ± 0.0	66 ± 0.1

https://doi.org/10.1371/journal.pone.0267714.t006

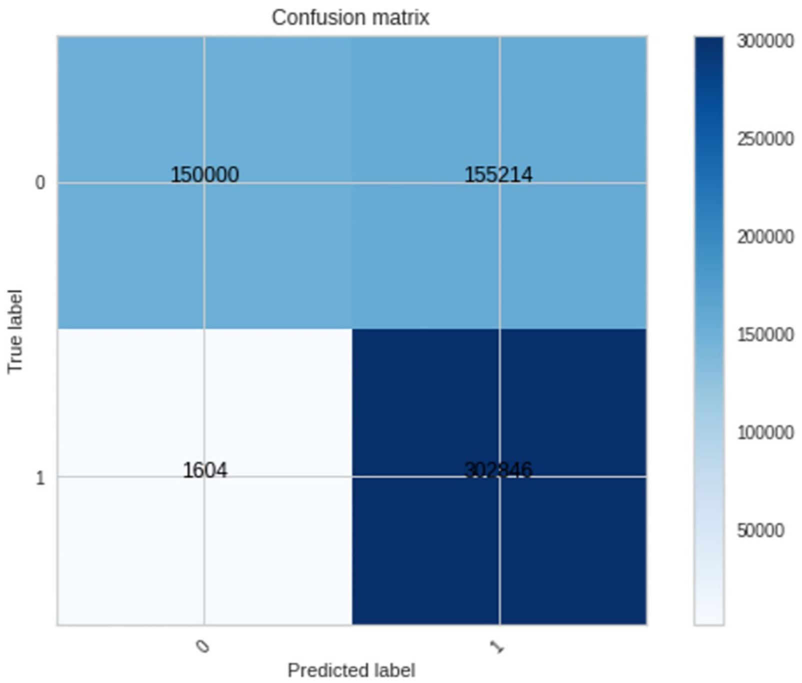


Fig 9. Confusion matrix of the XGBoost model for the *APC* gene.

https://doi.org/10.1371/journal.pone.0267714.g009

previous section, the models struggled to predict the Gleason grade group given DNA sequences and in this section of results, although there is plenty of room for improvement; the models were able to satisfactorily separate tumor DNA sequences from matched-normal DNA sequences. The confusion matrix of the highest performing model (XGBoost) is shown (Fig 9).

The main limitations of this work include the use of a small sample size, particularly the *BRCA 1* and *BRCA 2* DNA sequences. For this reason, the machine learning models were not able to competently distinguish Gleason grade group of 5 DNA sequences from Gleason grade group of 1 DNA sequences. The other limitation in this work include the lack of sufficient prior research on this topic, particularly research that has used DNA sequences as the only input source to machine learning or deep learning classifiers in the prediction of the Gleason grade group problem. Subsequently, it was difficult to benchmark the results of this work with those in the literature.

5 Conclusion

The goal of this work was to apply machine learning algorithms in the prediction of the Gleason grade group in blood DNA sequences of high-risk and low-risk prostate cancer patients. The machine learning models were not able to sufficiently discriminate between Gleason grade group of 5 DNA sequences from Gleason grade group of 1 DNA sequences. The reasons for this occurred as a result of having a large number of sequences that share a substantial amount of sequence homology. Even though this was circumvented by removing highly similar sequences, it was still not sufficient as the machine learning classifiers still produced a high number of false positives and a negligible amount of true negatives. Since the machine learning models were not able to discriminate between the DNA sequences of the two Gleason grade groups, they were further evaluated to determine their usefulness in the prediction of tumor

DNA sequences from matched-normal DNA sequences. In this new problem, the models performed acceptably better than before.

The future work involves the design of better data representation techniques that are sensitive enough to discover discriminatory signatures in small sample sizes of DNA sequences. These techniques should be generic in that they should not only be sensitive towards Gleason grade groups, but should extend to other prediction problems that are important in machine learning and cancer research.

Acknowledgments

The authors would like to thank the patients for their valued contributions, as well the Southern African Prostate Cancer Study (SAPCS), for providing the data used in the study. The authors would also like to thank fellow colleagues that reviewed this manuscript.

Author Contributions

Conceptualization: Mpho Mokoatle, Darlington Mapiye.

Data curation: Mpho Mokoatle, Vanessa M. Hayes, Riana Bornman.

Formal analysis: Mpho Mokoatle, Darlington Mapiye.

Investigation: Mpho Mokoatle.

Methodology: Mpho Mokoatle, Darlington Mapiye.

Resources: Mpho Mokoatle.

Supervision: Vukosi Marivate.

Validation: Mpho Mokoatle, Vanessa M. Hayes, Riana Bornman.

Visualization: Mpho Mokoatle.

Writing – original draft: Mpho Mokoatle.

Writing – review & editing: Mpho Mokoatle, Vukosi Marivate.

References

1. Cassim N, Ahmad A, Wadee R, Rebbeck T, Glencross D, George J. Prostate cancer age-standardised incidence increase between 2006 and 2016 in Gauteng Province, South Africa: A laboratory data-based analysis. South African Medical Journal. 2021; 111(1):26–32. https://doi.org/10.7196/SAMJ.2020.v111i1.14850
2. Pienta KJ, Esper PS. Risk factors for prostate cancer. Annals of internal medicine. 1993; 118(10):793–803. https://doi.org/10.7326/0003-4819-118-10-199305150-00007 PMID: 8470854
3. Heidenreich A, Bastian PJ, Bellmunt J, Bolla M, Joniau S, Mason M, et al. Guidelines on prostate cancer. European association of urology. 2012; p. 45. PMID: 22699678
4. Gann PH. Risk factors for prostate cancer. Reviews in urology. 2002; 4(Suppl 5):S3. PMID: 16986064
5. van der Leest M, Cornel E, Israël B, Hendriks R, Padhani AR, Hoogenboom M, et al. Head-to-head comparison of transrectal ultrasound-guided prostate biopsy versus multiparametric prostate resonance imaging with subsequent magnetic resonance-guided biopsy in biopsy-naive men with elevated prostate-specific antigen: a large prospective multicenter clinical study. European urology. 2019; 75 (4):570–578. PMID: 30477981
6. ANATOMY AND FUNCTION OF THE PROSTATE; 2019[Online]. Available from: https://prostate-ca.co.za/information-on-the-diagnosis-treatment-of-prostate-cancer/.
7. Togo Y, Yamamoto S. Prevention of infectious complications after prostate biopsy procedure. International Journal of Urology. 2017; 24(7):486–492. https://doi.org/10.1111/iju.13369 PMID: 28556409

8. Egevad L, Delahunt B, Evans AJ, Grignon DJ, Kench JG, Kristiansen G, et al. International Society of Urological Pathology (ISUP) grading of prostate cancer. The American journal of surgical pathology. 2016; 40(6):858–861. https://doi.org/10.1097/PAS.0000000000000642 PMID: 27023325

9. Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, et al. A contemporary prostate cancer grading system: a validated alternative to the Gleason score. European urology. 2016; 69(3):428–435. https://doi.org/10.1016/j.eururo.2015.06.046 PMID: 26166626

10. Özdamar Ş, Sarikaya Ş, Yildiz L, Atilla M, Kandemir B, Yildiz S. Intraobserver and interobserver reproducibility of WHO and Gleason histologic grading systems in prostatic adenocarcinomas. International urology and nephrology. 1996; 28(1):73–77. https://doi.org/10.1007/BF02550141 PMID: 8738623

11. Melia J, Moseley R, Ball R, Griffiths D, Grigor K, Harnden P, et al. A UK-based investigation of inter-and intra-observer reproducibility of Gleason grading of prostatic biopsies. Histopathology. 2006; 48 (6):644–654. https://doi.org/10.1111/j.1365-2559.2006.02393.x PMID: 16681679

12. Egevad L, Ahmad AS, Algaba F, Berney DM, Boccon-Gibod L, Compérat E, et al. Standardization of Gleason grading among 337 European pathologists. Histopathology. 2013; 62(2):247–256. https://doi.org/10.1111/his.12008 PMID: 23240715

13. Abdollahi A, Meysamie A, Sheikhbahaei S, Ahmadi A, Moradi-Tabriz H, Bakhshandeh M, et al. Inter/intra-observer reproducibility of Gleason scoring in prostate adenocarcinoma in Iranian pathologists. Urology journal. 2012; 9(2):486–490. PMID: 22641492

14. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB, et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. Human pathology. 2001; 32 (1):74–80. https://doi.org/10.1053/hupa.2001.21135

15. Veloso SG, Lima MF, Salles PG, Berenstein CK, Scalon JD, Bambirra EA. Interobserver agreement of Gleason score and modified Gleason score in needle biopsy and in surgical specimen of prostate cancer. International braz j urol. 2007; 33:639–651. https://doi.org/10.1590/S1677-55382007000500005 PMID: 17980061

16. Bottke D, Golz R, Störkel S, Hinke A, Siegmann A, Hertle L, et al. Phase 3 study of adjuvant radiotherapy versus wait and see in pT3 prostate cancer: impact of pathology review on analysis. European urology. 2013; 64(2):193–198. https://doi.org/10.1016/j.eururo.2013.03.029 PMID: 23522911

17. van der Kwast TH, Collette L, Van Poppel H, Van Cangh P, Vekemans K, DaPozzo L, et al. Impact of pathology review of stage and margin status of radical prostatectomy specimens (EORTC trial 22911). Virchows Archiv. 2006; 449(4):428–434. https://doi.org/10.1007/s00428-006-0254-x PMID: 16941153

18. Kvåle R, Møller B, Wahlqvist R, Fosså SD, Berner A, Busch C, et al. Concordance between Gleason scores of needle biopsies and radical prostatectomy specimens: a population-based study. BJU international. 2009; 103(12):1647–1654. https://doi.org/10.1111/j.1464-410X.2008.08255.x PMID: 19154461

19. Montironi R, Lopez-Beltran A, Cheng L, Montorsi F, Scarpelli M. Central prostate pathology review: should it be mandatory. Eur Urol. 2013; 64(2):199–201. https://doi.org/10.1016/j.eururo.2013.04.002 PMID: 23608669

20. Wulczyn E, Nagpal K, Symonds M, Moran M, Plass M, Reihs R, et al. Predicting prostate cancer specific-mortality with artificial intelligence-based Gleason grading. Communications Medicine. 2021; 1 (1):1–8. https://doi.org/10.1038/s43856-021-00005-3

21. Agalliu I, Gern R, Leanza S, Burk RD. Associations of high-grade prostate cancer with BRCA1 and BRCA2 founder mutations. Clinical Cancer Research. 2009; 15(3):1112–1120. https://doi.org/10.1158/1078-0432.CCR-08-1822 PMID: 19188187

22. Easton D, Thompson D, McGuffog L, Haites N, Schofield A, Scott R, et al. Cancer risks in BRCA2 mutation carriers. The breast cancer linkage consortium J Natl Cancer Inst. 1999; 91(15):1310–1316.

23. Johannsson O, Loman N, Möller T, Kristoffersson U, Borg Å, Olsson H. Incidence of malignant tumours in relatives of BRCA1 and BRCA2 germline mutation carriers. European journal of cancer. 1999; 35 (8):1248–1257. https://doi.org/10.1016/S0959-8049(99)00135-5 PMID: 10615237

24. Castro E, Eeles R. The role of BRCA1 and BRCA2 in prostate cancer. Asian journal of andrology. 2012; 14(3):409. https://doi.org/10.1038/aja.2011.150 PMID: 22522501

25. Nyberg T, Frost D, Barrowdale D, Evans DG, Bancroft E, Adlard J, et al. Prostate cancer risks for male BRCA1 and BRCA2 mutation carriers: A prospective cohort study. European urology. 2020; 77(1):24–35. https://doi.org/10.1016/j.eururo.2019.08.025 PMID: 31495749

26. Arvaniti E, Fricker KS, Moret M, Rupp N, Hermanns T, Fankhauser C, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Scientific reports. 2018; 8(1):1–11. https://doi.org/10.1038/s41598-018-30535-1 PMID: 30104757

27. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. The Lancet Oncology. 2020; 21(2):222–232. https://doi.org/10.1016/S1470-2045(19)30738-7 PMID: 31926806

28. Chaddad A, Kucharczyk MJ, Desrosiers C, Okuwobi IP, Katib Y, Zhang M, et al. Deep radiomic analysis to predict gleason score in prostate cancer. IEEE Access. 2020; 8:167767–167778. https://doi.org/10.1109/ACCESS.2020.3023902

29. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology. 2020; 21(2):233–241. https://doi.org/10.1016/S1470-2045(19)30739-9 PMID: 31926805

30. Lucas M, Jansen I, Savci-Heijink CD, Meijer SL, de Boer OJ, van Leeuwen TG, et al. Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. Virchows Archiv. 2019; 475(1):77–83. https://doi.org/10.1007/s00428-019-02577-x PMID: 31098801

31. Liu H, Tang K, Peng E, Wang L, Xia D, Chen Z. Predicting Prostate Cancer Upgrading of Biopsy Gleason Grade Group at Radical Prostatectomy Using Machine Learning-Assisted Decision-Support Models. Cancer Management and Research. 2020; 12:13099. https://doi.org/10.2147/CMAR.S286167 PMID: 33376402

32. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.

33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

34. Fletez-Brant C, Lee D, McCallion AS, Beer MA. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. Nucleic acids research. 2013; 41(W1):W544–W556. https://doi.org/10.1093/nar/gkt519 PMID: 23771147

35. Bafna P, Pramod D, Vaidya A. Document clustering: TF-IDF approach. In: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). IEEE; 2016. p. 61–66.

36. Qaiser S, Ali R. Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications. 2018; 181(1):25–29. https://doi.org/10.5120/ijca2018917395

37. Shahmirzadi O, Lugowski A, Younge K. Text similarity in vector space models: a comparative study. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). IEEE; 2019. p. 659–666.

38. Bro R, Smilde AK. Principal component analysis. Analytical methods. 2014; 6(9):2812–2831. https://doi.org/10.1039/C3AY41907J

39. Ng P. dna2vec: Consistent vector representations of variable-length k-mers. arXiv preprint arXiv:170106279. 2017.

40. Zou Q, Xing P, Wei L, Liu B. Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. Rna. 2019; 25(2):205–218. https://doi.org/10.1261/rna.069112.118 PMID: 30425123

41. Asgari E, Mofrad MR. Continuous distributed representation of biological sequences for deep proteomics and genomics. PloS one. 2015; 10(11):e0141287. https://doi.org/10.1371/journal.pone.0141287 PMID: 26555596

42. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013.

43. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p. 785–794.

44. Elavarasan D, Vincent DR. Reinforced XGBoost machine learning model for sustainable intelligent agrarian applications. Journal of Intelligent & Fuzzy Systems. 2020;(Preprint):1–16.

45. Yang S, Yu X, Zhou Y. Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In: 2020 International workshop on electronic communication and artificial intelligence (IWECAI). IEEE; 2020. p. 98–101.

46. Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. In: Thirteenth annual conference of the international speech communication association; 2012.

47. Zhao Z, Chen W, Wu X, Chen PC, Liu J. LSTM network: a deep learning approach for short-term traffic forecast. IET Intelligent Transport Systems. 2017; 11(2):68–75. https://doi.org/10.1049/iet-its.2016.0208

48. Smagulova K, James AP. A survey on LSTM memristive neural network architectures and applications. The European Physical Journal Special Topics. 2019; 228(10):2313–2324. https://doi.org/10.1140/epjst/e2019-900046-x

49. Dey R, Salem FM. Gate-variants of gated recurrent unit (GRU) neural networks. In: 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS). IEEE; 2017. p. 1597–1600.

50. Wang Y, Liao W, Chang Y. Gated recurrent unit network-based short-term photovoltaic forecasting. Energies. 2018; 11(8):2163. https://doi.org/10.3390/en11082163

51. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555. 2014.

52. Xu B, Guo X, Ye Y, Cheng J. An Improved Random Forest Classifier for Text Categorization. J Comput. 2012; 7(12):2913–2920.

53. Radovanović S, Ivić M. Enabling Equal Opportunity in Logistic Regression Algorithm. Management: Journal of Sustainable Business and Management Solutions in Emerging Economies. 2021.

54. Pratiwi H, Windarto AP, Susliansyah S, Aria RR, Susilowati S, Rahayu LK, et al. Sigmoid Activation Function in Selecting the Best Model of Artificial Neural Networks. In: Journal of Physics: Conference Series. vol. 1471. IOP Publishing; 2020. p. 012010.

55. Alin A. Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics. 2010; 2(3):370–374. https://doi.org/10.1002/wics.84

56. Hu G, Kurgan L. Sequence similarity searching. Current protocols in protein science. 2019; 95(1):e71. https://doi.org/10.1002/cpps.71 PMID: 30102464

57. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

58. Holman CM. Protein Similarity Score: A Simplified Version of the BLAST Score as a Superior Alternative to Percent Identity for Claiming Genuses of Related Protein Sequences. Santa Clara Computer & High Tech LJ. 2004; 21:55.

59. RNN vs GRU vs LSTM; 2020 [Online]. Available from: https://medium.com/analytics-vidhya/rnnvs-gru-vs-lstm-863b0b7b1573.