KeAi
CHINESE ROOTS
GLOBAL IMPACT

Contents lists available at ScienceDirect

Data Science and Management

journal homepage: www.keaipublishing.com/en/journals/data-science-and-management

Research article

# Application of support vector machine algorithm for early differential diagnosis of prostate cancer

Boluwaji A. Akinnuwesi [a,*], Kehinde A. Olayanju [b], Benjamin S. Aribisala [c], Stephen G. Fashoto [a], Elliot Mbunge [a], Moses Okpeku [d], Patrick Owate [c]

[a] *Department of Computer Science, Faculty of Science and Engineering, University of Eswatini, Kwaluseni, M201, Swaziland*
[b] *Department of Computer Science Education, Federal College of Education (Technology), Akoka, Lagos State, 100213, Nigeria*
[c] *Department of Computer Science, Faculty of Science, Lagos State University, Ojo, Lagos State, 102101, Nigeria*
[d] *Department of Genetics, University of KwaZulu-Natal, Durban, 4041, South Africa*

ARTICLE INFO

*Keywords:*
Confusable diseases
Computational intelligence
Early differential diagnosis
Logistic regression
Prostate cancer
Support vector machine

ABSTRACT

Prostate cancer (PCa) symptoms are commonly confused with benign prostate hyperplasia (BPH), particularly in the early stages due to similarities between symptoms, and in some instances, underdiagnoses. Clinical methods have been utilized to diagnose PCa; however, at the full-blown stage, clinical methods usually present high risks of complicated side effects. Therefore, we proposed the use of support vector machine for early differential diagnosis of PCa (SVM-PCa-EDD). SVM was used to classify persons with and without PCa. We used the PCa dataset from the Kaggle Healthcare repository to develop and validate SVM model for classification. The PCa dataset consisted of 250 features and one class of features. Attributes considered in this study were age, body mass index (BMI), race, family history, obesity, trouble urinating, urine stream force, blood in semen, bone pain, and erectile dysfunction. The SVM-PCa-EDD was used for preprocessing the PCa dataset, specifically dealing with class imbalance, and for dimensionality reduction. After eliminating class imbalance, the area under the receiver operating characteristic (ROC) curve (AUC) of the logistic regression (LR) model trained with the downsampled dataset was 58.4%, whereas that of the AUC-ROC of LR trained with the class imbalance dataset was 54.3%. The SVM-PCa-EDD achieved 90% accuracy, 80% sensitivity, and 80% specificity. The validation of SVM-PCa-EDD using random forest and LR showed that SVM-PCa-EDD performed better in early differential diagnosis of PCa. The proposed model can assist medical experts in early diagnosis of PCa, particularly in resource-constrained healthcare settings and making further recommendations for PCa testing and treatment.

## 1. Introduction

Cancer is a chronic and noncommunicable disease that remains a significant global public health problem. Cancer mortality is projected to increase to 11 million deaths annually by 2030, with the majority of these occurring in regions of the world with the least capacity to respond. The United Nations General Assembly Special Session in 2011 codified the key aspects of cancer prevention and control, which includes intensifying primary prevention through diagnosis, treatment, and care (Cannon et al., 2012). However, this is undermined by the impact of social, economic, and environmental determinants of health, such as poverty, illiteracy, gender inequality, social isolation, stigma, socio-economic status, poor access to health facilities, and early diagnosis equipment, particularly in developing countries. Among the various types of cancer, prostate cancer (PCa) is the third leading cause of cancer-related mortality in men globally (Brabletz et al., 2018; Rahman and Chowdhury, 2016; Steinestel et al., 2019). It is becoming more prevalent worldwide, with an estimated population of over 1.4 million new cases and over 370,000 deaths in 2020 (Sung et al., 2021; WHO, 2020).

The prostate is a component of the male reproductive system, located in the pelvis below the urinary bladder and front of the rectum, as shown in Fig. 1. It surrounds part of the urethra and typically measures approximately 3 cm in length and 20 g in weight in an adult man. The human prostate is a zinc-accumulating and citrate-producing organ that helps to produce and store seminal fluid (Cunha et al., 1987). Prostate

Peer review under responsibility of Xi'an Jiaotong University.
[*] Corresponding author.
*E-mail address:* bakinnuwesi@uniswa.sz (B.A. Akinnuwesi).

https://doi.org/10.1016/j.dsm.2022.10.001
Received 17 May 2022; Received in revised form 6 October 2022; Accepted 7 October 2022
Available online 19 October 2022

glands produce approximately 20% of seminal fluid, and diseased prostrates affect urination, ejaculation, and defecation.

The symptoms of PCa are often similar to those of other diseases, particularly in the early stages. The signs and symptoms of PCa include difficulty in urinating, pelvic pain, bloody urine, and fatigue due to a deficiency of red blood cells. PCa is associated with risk factors, such as old age, heredity, and race. This implies that having an immediate member of the family infected with PCa increases the risk because it is hereditary (Brabletz et al., 2018; Hagiwara et al., 2018). In addition, some behavioral and dietary risk factors are associated with PCa, including high consumption of milk products, processed meat, or diets low in certain vegetables (Alexander et al., 2010; Bylsma and Alexander, 2015). A few surveys have found an unusual association between PCa and gonorrhea (Caini et al., 2014).

The prevalence of PCa among men has been a global concern (Houston et al., 2018). For example, in the U.S. approximately 191,000 new cases were estimated in 2020, with 33, 330 deaths due to prostate cancer. Older men are a susceptible group, and, based on a report from the American Cancer Society (ACS) (American Cancer Society, 2022), one of every nine men worldwide is diagnosed with PCa in their lifetime. Moreover, the ACS report also stated that approximately six out of ten cases of PCa are diagnosed in men aged 65 years and above, whereas PCa is rare in men aged below 40 years old. This aligns with the findings reported by Zhou et al. (2016) on PCa incidences worldwide. In addition, a world health organization (WHO) report showed that PCa led to the deaths of over 12,000 males in Nigeria (WHO, 2020). Similarly, PCa ranked first among cancer diseases in Nigeria with high mortality, and the death rate recorded by WHO is 46.41 per 100,000 population. This ranked PCa in Nigeria as the 12th leading cause of cancer mortality in men globally. Therefore, there is a need to develop feasible computational intelligence-based tools to improve primary PCa prevention, diagnosis, prognosis, and treatment. This study aimed to develop a feasible machine-learning-based PCa early differential diagnosis model that can be used in resource-constrained healthcare settings for the early detection and diagnosis of PCa to reduce mortality.

Herein, we present a support vector machine model for early differential diagnosis of PCa (SVM-PCa-EDD). SVM-PCa-EDD is a computational approach based on SVM techniques to classify persons with and without PCa using a dataset from the Kaggle Healthcare repository (Kaggle, 2022). SVM, as a supervised learning algorithm (Cortes and Vapnik, 1995), has demonstrated high performance in solving classification problems in many biological and medical fields, including bioinformatics (Ng and Mishra, 2007; Rice et al., 2005). An SVM algorithm discriminates between two classes by generating a hyperplane that optimally separates classes after input data have been mathematically transformed into a high-dimensional space. It is data-driven and model-free; therefore, it has important discriminative power for classification, particularly in cases where sample sizes are small and many variables are involved.

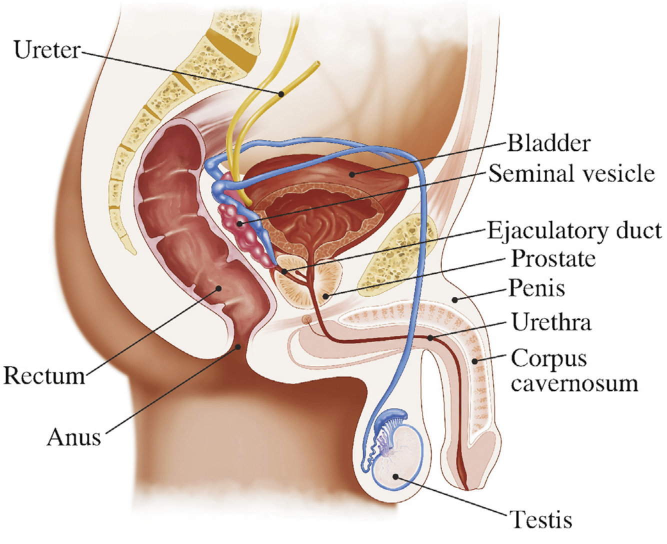

**Fig. 1.** Prostate physiology.

However, deep learning models, such as convolutional neural networks (CNN), multilayer perceptron (MLP), radial basis function network (RBFN), recurrent radial basis function (RRBF), recurrent neural networks (RNNs), long short-term memory networks (LSTMs), restricted Boltzmann machines (RBMs), generative adversarial networks (GANs), and deep belief networks (DBNs), were not considered in this study because our dataset was not related to cancer image data, cancer image recognition, brain imaging, pattern recognition and prediction, time series or prediction, and exploratory data analysis of high-dimensional datasets. We utilized a dataset with PCa symptoms because our research focuses on early diagnosis of PCa before it develops to the formation of cancer lumps that require image processing techniques. Therefore, our interest is in a dataset of symptoms exhibited in patients before the formation of PCa lumps. The symptoms include difficulty in urinating (slow or weak stream), frequent urination (or dysuria), especially throughout the night, urinary incontinence, varying from slight to complete loss of bladder control, blood in urine, erectile dysfunction, painful urination, and pain in the hips, lower back, and chest. We employed SVM because it is widely used to solve data-classification and outlier-detection problems, data preprocessing, class imbalance, and dimensionality reduction.

Our model (SVM-PCa-EDD) was used to classify patients with and without PCa. We used the PCa dataset from the Kaggle Healthcare repository to develop and validate the SVM model for a classification scheme of diagnosed PCa vs. no PCa. The PCa dataset was a labelled dataset consisting of 250 features and one class of features. The patient data considered were general background information, serum information, and the surface-enhanced laser desorption ionization (SELDI) lipid profile. The background information of patients included age, body mass index (BMI), race, family history, difficulty in urinating, weak urine stream, blood in semen, bone pain, and erectile dysfunction. The collected data were preprocessed to remove class imbalance, incompleteness, noise, and other inconsistencies. The proposed SVM-PCa-EDD was used to pre-process the PCa dataset, to remove class imbalance and reduce dimensionality. The performance of SVM-PCa-EDD was evaluated to determine its accuracy, specificity, and sensitivity.

## 2. Review of related works

### 2.1. Overview of PCa

PCa was first discovered in 1853 (Adams, 1853). It was a rare disease, and its detection methods were poor in the 19th century. Cancers emerge from an ongoing Darwinian evolutionary process, which often leads to multiple subclones within a single primary tumor (Gerlinger et al., 2012). This process leads to the formation of metastases which are a major cause of morbidity in cancer cases. Metastasis is the spread of cancer cells from the place where they first formed to other parts of the body. In PCa, prostate gland cells mutate into cancer cells. Men who present with metastatic PCa receive primary androgen deprivation therapy (ADT) to which they naturally develop resistance (De Bono et al., 2011). Many scholars have identified recurrent somatic mutations, copy number alterations, and oncogenic structural DNA rearrangements in primary PCa (Kote-Jarai et al., 2011; Rahman and Chowdhury, 2016; Zhang et al., 2016), and PCa is linked to urinary dysfunction because the prostate gland is situated immediately above the proximal part of the urethra (which is therefore called the prostatic urethra). Seminal fluids are deposited in the prostatic urethra by the vas deferens, and PCa is implicated in failure of erection and painful ejaculation (Zhang et al., 2016). Other symptoms include pain in bones, vertebrae, pelvis, and ribs.

PCa in the bones of the spine can compress the spinal cord and thereby cause leg weakness, tingling, and urinary and fecal incontinence.

Initial PCa treatments were based on surgeries to relieve urinary obstruction, including radical perineal prostatectomy and orchiectomy, but had limited success (Adams, 1853). In the 20th century, transurethral resection of the prostate (TURP) replaced radical prostatectomy for symptomatic relief of obstruction because it could better preserve erectile function. In 1983, radical retropubic prostatectomy (RP) was developed to treat PCa (Yaxley et al., 2016). This surgical approach allows the removal of prostate and lymph nodes while leaving the penile function intact. Recently, several PCa clinical treatments have been developed, including the use of estrogen to reduce the production of testosterone in men with metastatic PCa (Denmeade and Isaacs, 2002) and radiation therapy, such as external beam radiotherapy and brachytherapy with implanted radioactive metallic seeds for PCa (Denmeade and Isaacs, 2002).

Some forms of clinical diagnostic methods available for diagnosing PCa are presented below.

### 2.1.1. Prostate imaging

Ultrasound and magnetic resonance imaging (MRI) are imaging methods used for PCa diagnosis. Transrectal ultrasound is used for ultrasound-guided needle prostate biopsy by creating an image of organs in the pelvis, and most often reveals a hypoechoic lesions in PCa conditions (Schoots et al., 2015). However, research has shown that MRI has a better soft tissue resolution than ultrasound (Bonekamp et al., 2011). The role of an MRI differs depending on the risk level of PCa diagnosis. In low-risk PCa, MRI can be used for active surveillance of the patient, whereas in high-risk PCa, it helps in detecting bone disease (Barentsz et al., 2012). The fusion of MRI with ultrasound is often used to identify targets for prostate biopsy (Natarajan et al., 2011). Prostate MRI is used for surgical planning for patients undergoing robotic prostatectomies. It helps surgeons decide on the following: resection or sparing the neurovascular bundle, determination of a regain of urinary continence, and assessment of surgical difficulty.

### 2.1.2. Biopsy

Biopsies can be performed if a cancer is suspected. Outpatients routinely undergo prostate biopsy procedures and rarely require hospitalization. Antibiotics are usually administered to prevent complications such as sepsis, urinary tract infections, and fever; however, some discomfort is inevitable (Yaghi and Kehinde, 2015). Biopsies can be standard or targeted, with the latter being more common. Targeted MRI/ultrasound was found to accurately detect early stages of high-risk cancer, and significantly improve the detection of low-risk cancer. However, targeted biopsies may have clinical implications (Siddiqui et al., 2015). Synthesized evidence further revealed that hematuria and hematospermia are common in tested patients. Hospitalization is frequently necessary owing to complications after prostate biopsy. Therefore, it is generally recommended that biopsies include antimicrobial and pain-management drugs. Imaging biopsy is an effective method of diagnosing PCa. However, many patients undergo prostate biopsy attended with pain, stress, bleeding, and high cost, yet with only a small possibility of accurately detecting cancer (Taitt, 2018). This renders prostate biopsy imperfect and ineffective at the earlier stages of PCa diagnosis.

### 2.1.3. Digital rectal examination (DRE)

DRE is a physical examination of the lower rectum, pelvis, and lower belly by a physician, and is used to check health problems, including checking the prostate. In a primary healthcare setting, DRE is considered a routine screening for PCa (Naji et al., 2018). Doctors recommend further testing if there are abnormalities in the shape, texture, and size of the gland. The research reported by Naji et al. (2018) did not recommend the use of DRE for routine screening for PCa in primary care to minimize diagnostic testing, overdiagnosis, and overtreatment. Moreover, there is insufficient evidence to support the efficacy of DRE.

### 2.1.4. Prostate-specific antigen (PSA)

The cells of the prostate gland produce PSA, and its level in the blood of male subjects is measured using a PSA test. A higher than normal PSA level in the bloodstream is an indicator for prostate infection, inflammation, enlargement, or cancer (National Cancer Institute, 2018). The PSA test has been recognized to help in detecting early stages of PCa. However, it potentially has the following consequences (National Cancer Institute, 2018): it could lead to overtreatment, which can expose patients unnecessarily to complications and harmful side effects of treatments (e.g., surgery and radiation therapy) of early PCa, such as urinary incontinence, bowel malfunction, erectile dysfunction, and infection. The PSA test may also generate inaccurate results and consequently create anxiety if it is false positive or false assurance of no PCa occurrence if it is false negative.

The clinical methods for diagnosing PCa are partially sufficient for its detection. However, these methods are used only in the full-blown stage of PCa. PCa is not detectable by imaging or discovered in bioptic extracts in its early stages, and the same applies to the use of DRE. Although PSA testing supports early detection, the results are not always accurate. Therefore, some doctors argue against its use (National Cancer Institute, 2018). Therefore, none of these methods can be used to accurately detect the possibility of PCa in a patient prior to the development of PCa lumps, when the patient exhibits PCa-related symptoms that could be mistaken for some other diseases. Therefore, we need to develop a computational system using a machine learning technique that could help in the early differential diagnosis of PCa before it begins to grow. Such a computational system serves as a decision support tool for physicians in the early screening of patients for PCa, without clinical complications that arise from using biopsies, DREs, and PSA tests. Differential diagnosis tools can assist health experts in differentiating diseases with similar symptoms and determining the percentage of overlap between mixed symptoms (Akinnuwesi et al., 2020; Uzoka et al., 2016). This helps doctors to ensure accurate diagnosis and thereby recommend the correct treatment for patients. For instance, PCa exhibits symptoms similar to those of benign prostate hyperplasia (BPH) (De Vincentis et al., 2018; Taitt, 2018). Early differential diagnosis tools have become imperative for categorically detecting overlapping symptoms to enable early detection and prevent misdiagnoses (Li et al., 2018a; Pereira et al., 2020).

PCa mortality in men is becoming a grave concern; therefore, more efficient and effective diagnostic and treatment approaches are required to minimize it. Moreover, PCa symptoms are confusable with other prostrate diseases, such as BPH, making it difficult to diagnose PCa at an early stage using existing clinical methods. The symptoms are not as obvious in the early stages as they are in later stages. Meanwhile, delayed diagnosis may result in complications and metastases, which can eventually result in death (Smith et al., 2015).

## 2.2. Risk factors associated with PCa

PCa is mostly associated with risk factors, such as age, obesity, and family history. However, risk factors, such as BPH, smoking, and alcohol consumption, could make low-risk patients vulnerable to PCa (Gann, 2002). The incidence of PCa is rare in men below the age of 40 but becomes more common as men advance in age. Autopsy studies on diseased Chinese, Israeli, German, Jamaican, Swedish, and Ugandan men found PCa in 30% of men in their fifties and 80% of men in their seventies (Tian and Osawa, 2015). Genetic factors are considered among PCa risk factors, and lack of exercise or a sedentary lifestyle may also contribute to an extent (Marley and Nan, 2016). Obesity and elevated blood testosterone levels have also been identified as causative factors of PCa (Kumagai et al., 2015; Moyad, 2002; Parikesit et al., 2016).

Consumption of fruits and certain vegetables has been acknowledged to limit the risk of developing PCa. Higher meat consumption may mean a higher risk, as stated by Wolk (2017). However, lower blood levels of vitamin D and the use of cholesterol-lowering medications may increase the risk of developing PCa (Murtola et al., 2018; Platz et al., 2009).

Similarly, the risk of PCa is elevated when there are infections that could cause inflammation of the prostate glands. Studies have also shown that environmental factors could contribute to PCa occurrence (Tse et al., 2017; Vaidyanathan et al., 2017).

### 2.3. Differential diagnosis

Differential diagnosis is a process that helps differentiate between diseases having similar symptoms and risk factors. It is a systematic diagnostic process performed on patients to accurately diagnose a disease that shares symptoms with related diseases and thrives under identical conditions (Mann, 1990; Sand, 2015; Uzoka et al., 2016). For example, diseases such as HIV/AIDS, malaria, flu, tuberculosis, COVID-19, Ebola virus disease, and cholera have similar symptoms. Therefore, when patients exhibit one or more of these symptoms, physicians need to subject them to a differential diagnostic process to establish the actual one among the multiple related diseases. Symptoms of BPH, such as trouble starting a urine stream or making more than a dribble, frequent urination, especially at night, feeling that the bladder has not fully emptied, weak or slow urine stream, are similar to those of PCa, and patients that experience any of these symptoms are expected to undergo a differential diagnosis procedure to ascertain the exact disease (i.e., BPH or PCa) that they are afflicted with. Therefore, a differential diagnosis process involves weighing the probability of a disease against the probabilities of other related diseases that may account for a patient's illness.

Computational algorithms, such as soft computing and machine learning, have been used for differential diagnosis using details including symptoms, patient history, and medical knowledge to guide the process and ensure accurate diagnosis (Liberman et al., 2016).

Differential diagnosis involves the following four steps (Federman and Chanko, 2007): (1) collecting information about the patients and creating a list of symptoms; (2) listing possible causes of the symptoms; (3) prioritizing the list by placing the most dangerous possible causes of the symptoms at the top of the log; (4) eliminating or establishing the possible causes, starting with the most dangerous condition and working down the list.

### 2.4. Overview of computational intelligence techniques applied in PCa diagnosis

In this section, we discuss various intelligent computational techniques or models that have been developed for PCa detection, diagnosis, and prediction. Intelligent computational techniques seek to combine evolutionary algorithms with machine learning algorithms to optimize the predictive performance of the model (Mumford and Jain, 2009). Machine learning algorithms are effective in predictive modelling, wherein components are developed to learn from existing patient historical data and make predictions on new patient data (Kakade et al., 2009). Computational intelligence approaches also help to process cases of imprecision and uncertainty through which clinical and biological data could be vague or ambiguous (Cosma et al., 2017). For example, computational intelligent algorithms have been applied to risk prediction models for breast cancer (Turner et al., 2007), cardiovascular disease (Nair and Vijaya, 2010), and lung cancer. Similarly, various intelligent computational techniques have been adopted for PCa prognosis, vis-à-vis diagnosis. These techniques include artificial neural networks (ANN), genetic algorithms (GA), evolutionary algorithms (EA), fuzzy systems, and SVM.

A survey of computational intelligence methods that have been applied to PCa prediction was presented by Cosma et al. (2017). The techniques considered in the study were ANN, fuzzy-based techniques, metaheuristic optimization algorithms, deep learning, Markov models, and Bayesian-based techniques. Artificial immune network, ant colony optimization, and particle swarm optimization were identified by the authors as techniques that have been most used for optimizing the performance of predictive models for PCa. Similarly, Goldenberg et al. (2019) emphasized the strong need for the application of artificial intelligence and machine learning techniques for the proper management of PCa.

Computational intelligence techniques have also been applied to prostate segmentation. Tian et al. (2018) proposed a deep CNN model that automatically segments the prostate. Similarly, as was described by To et al. (2018), a 3D CNN model was used to segment the prostate in magnetic resonance (MR) images. A CNN-based algorithm was demonstrated by Clark et al. (2017) to delineate the transition zone (TZ) of the prostate gland in diffusion-weighted imaging (DWI). The performance results showed an average accuracy of 0.97 for the detection of image slices with and without the prostate gland. Similarly, a deep CNN model was used for prostate segmentation by Cheng (2017a, 2017b).

A deep CNN model was proposed by Kwak and Hewitt (2017) to diagnose PCa. The model was evaluated using two tissue microarrays (TMA), and the area under the receiver operating characteristic (ROC) curve (AUC) was 0.95, which demonstrated that the model can potentially improve PCa pathology. However, this model cannot be used for early differential diagnosis of PCa. Similarly, the region-based CNN (R-CNN) model for PCa diagnosis presented by Li et al. (2018b) helped to detect epithelial cells with an accuracy of 99.07% and an AUC of 0.998 and performed Gleason grading tasks with a mean intersection over union (IoU) of 79.56% and pixel accuracy of 89.40%; however, the R-CNN model cannot be used for early differential diagnosis of PCa. Silva-Rodríguez et al. (2020) proposed that a deep CNN-based automated system was proposed to help pathologists in the analysis of prostate whole-slide images. The analyses were as follows: Gleason grading of local structures, detection of cribriform patterns, and Gleason scoring of the entire biopsy. This work did not consider the combination of low- and high-level features for classification, and the detection of a cribriform pattern in the PCa cells was not included as a predictive factor in the end-to-end training of the deep-CNN model. Moreover, the model is not suitable for the early differential diagnosis of PCa. Furthermore, the automated system proposed by Lokhande et al. (2020) to grade prostate biopsies using deep Carcino-Net could also not be applied to early differential diagnosis of PCa. The deep CNN model proposed by Aldoj et al. (2020) was applied to the semi-automatic classification of PCa using multiparametric MR. The model helped to detect clinically significant PCa and is characterized by good AUC, sensitivity, and high specificity; however, it is not suitable for early differential diagnosis of PCa.

In summary, various scholars have developed and implemented different computational algorithms and methods to assist pathologists in the management of PCa in the following forms: (1) PCa diagnosis (Alkadi et al., 2019; Alkhateeb et al., 2020; Kott et al. 2021), (2) clinically significant PCa prediction (Bernatz et al., 2020), Gleason pattern prediction (Antonelli et al., 2019), PCa aggressiveness prediction (Liu et al., 2019), and (3) automatic Gleason grading of PCa (Bulten et al., 2020; Nir et al., 2019).

### 2.5. Deductions from literature review

Studies conducted on PCa diagnosis relied more on biopsy-based methods than on prostate imaging methods. In addition, most of the studies that relied on biopsy-based methods of diagnosis were restricted to manual diagnosis, while a few studies recommended ANN as a computational technique. Moreover, techniques such as deep CNN have been used for automatic segmentation of the prostate in MR images, detection of PCa, diagnosis of PCa, the Gleason system of grading of histological images, automatic daily analysis of prostate biopsies, automatic grading of prostate biopsies, and semi-automatic PCa classification. None of the studies focused on the early differential diagnosis of PCa by considering the early symptoms of PCa that are also present in other diseases. The following conclusions were drawn from the literature review:

(i) PCa is more likely in aging adults than in any other age groups.
(ii) Its effect and metastasis are more effectively prevented when diagnosed early.
(iii) PCa exhibits characteristics that are similar to those of other diseases, particularly BPH.
(iv) The features used to develop predictive models included age, PSA, weight, BMI, smoking habit, systolic and diastolic blood pressures, and Gleason score (i.e., grade given to PCa based on the arrangement of the cancer cells in the prostate; score assigned on a scale of 3–5 from two different locations).

## 3. Materials and methods

### 3.1. Analysis of perspectives of physicians on PCa diagnosis and treatment

We conducted a study to determine the perceptions of clinicians (urologists, oncologists, general surgeons, obstetricians, gynecologists, pediatricians, and non-specialist doctors) regarding early diagnosis and treatment of PCa in adults and children in Nigeria. Furthermore, we captured their views on the adoption and use of a computational intelligence system to complement their efforts in PCa diagnosis. We designed and randomly distributed questionnaires in twenty different hospitals in six local government areas in Lagos State, Nigeria. We conducted physical interviews with doctors (i.e., specialist and non-specialist doctors) using questionnaires. 60 questionnaires were administered to the doctors. 48 questionnaires (80%) were successfully completed and returned by doctors (i.e., specialist and non-specialist doctors). We successfully interacted with 10 specialist doctors and 38 non-specialist doctors across the hospitals. A dearth of specialist doctors was observed in the hospitals. We carried out the necessary descriptive statistics based on the collected data.

#### 3.1.1. Measures

The questionnaire was divided into two sections. The first section focused on the demographic, job description, and knowledge and experience of the doctors in the use of healthcare-related application software, and their involvement in healthcare software development in their hospitals. The second section of the questionnaire focused on the following major constructs: the level of computer literacy of the doctors, their adoption and use of intelligence-based applications for diagnosis and therapy, involvement in the development of medical computational systems and applications, patient reactions to the use of computational systems for diagnosis, and commitment of hospital management to the adoption of computational equipment and applications that can help complement the efforts of doctors. A 5-point Likert scale was used to enable the doctors to express their level of agreement or disagreement with most of the questions in the questionnaire. Microsoft Excel was used to analyze the data.

#### 3.1.2. Descriptive statistics of respondents' characteristics

Table 1 presents descriptive statistics based on the doctors' responses: 79.2% of the doctors were male and 20.8% were female. The majority of the doctors were over 40 years old (66.7%), which reflects the experience of doctors with over 11 years of work experience (64.6%). They had also diagnosed and treated patients infected with PCa, and they stated that it is common among adult males aged ≥50 years. This finding implies that PCa is a common disease in adult males. Among the specialist doctors and consultants, 50% were urologists and oncologists, 20% were obstetricians and gynecologists, 10% were pediatricians, and 20% were general surgeons; 40% had referred patients to consultant urologists and oncologists, and were aware of the risk factors and symptoms associated with PCa. In addition, 10.4% of the respondents had their highest qualification as doctorate degrees (Ph.D.) and were specialists and consultants; 25% had master's degrees (MMeds) as highest qualification, and 64.6% had

**Table 1**
Descriptive statistics of respondents' characteristics.

| Number | Variables | Description | Specialist doctors ($n$ = 10) | | Non-specialist doctors ($n$ = 38) | |
|---|---|---|---|---|---|---|
| | | | Frequency | Percentage (%) | Frequency | Percentage (%) |
| 1 | Age group | Below 25 years | 0 | 0 | 0 | 0 |
| | | 25–40 years | 1 | 10 | 15 | 39.5 |
| | | Above 40 years | 9 | 90 | 23 | 60.5 |
| 2 | Gender | Male | 7 | 70 | 31 | 81.6 |
| | | Female | 3 | 30 | 7 | 18.4 |
| 3 | Specialization | Urologist and oncologist | 5 | 50 | - | - |
| | | General surgeon | 2 | 20 | - | - |
| | | Obstetrician and gynecologist | 2 | 20 | - | - |
| | | Pediatrician | 1 | 10 | - | - |
| 4 | Highest qualification | Ph.D. | 5 | 50 | 0 | 0 |
| | | MMed | 4 | 40 | 8 | 21.1 |
| | | MBBS | 1 | 10 | 30 | 78.9 |
| 5 | Level in the hospital management | Top management | 8 | 80 | 0 | 0 |
| | | Mid-management | 2 | 20 | 10 | 26.3 |
| | | Operational | 0 | 0 | 28 | 73.7 |
| 6 | Years of work experience | Below 5 years | 0 | 0 | 10 | 26.3 |
| | | 6–10 years | 1 | 10 | 6 | 15.8 |
| | | 11–20 years | 3 | 30 | 10 | 26.3 |
| | | Above 20 years | 6 | 60 | 12 | 31.6 |
| 7 | Computer technology literacy level | Above average | 6 | 60 | 27 | 71.1 |
| | | Average | 4 | 40 | 11 | 28.9 |
| | | Not literate | 0 | 0 | 0 | 0 |
| 8 | Disease diagnosis method used | Use of computer medical application | 2 | 20 | 13 | 34.2 |
| | | Use of conventional clinical methods | 8 | 80 | 25 | 65.8 |
| 9 | Adoption and use of intelligent application for PCa diagnosis | Strongly agree | 8 | 80 | 28 | 73.7 |
| | | Agree | 2 | 20 | 10 | 26.3 |
| | | Indifferent | 0 | 0 | 0 | 0 |
| | | Disagree | 0 | 0 | 0 | 0 |
| 10 | Involvement in application development process for PCa diagnosis | Involved | 0 | 0 | 0 | 0 |
| | | Not Involved | 10 | 100 | 38 | 100 |

Note: $n$ represents sample population; Master of Medcine (MMed); Bachelor of Medcine and Bachelor of Surgery (MBBS).

their highest qualification as bachelor's degrees (MBBSs); 58.3% of the doctors operated at non-management levels in their hospitals, and 41.7% operated at the top- and mid-level management in their hospitals.

The findings of this study show that doctors are computer/digitally literate. They use computers (e.g., notebooks, laptops, or desktop computers), computer applications, Internet, smart phones, and electronic tablets for various purposes. Their computer literacy implies that they can easily understand, adopt, and use any intelligent computer system meant for PCa diagnosis. Despite a good computer literacy level, only 31.3% of the doctors used applications (apps) for diagnosis, with the others (68.7%) following conventional (clinical) diagnostic methods.

All the doctors (100%) agreed that an intelligent application for early diagnosis of PCa will be helpful, and are looking forward to such an application. None of the doctors had been involved in the development of intelligent applications for PCa diagnosis.

The doctors indicated that they encounter problems when diagnosing and managing patients with PCa. Therefore, if a PCa diagnosis application is developed, some of their challenges can be alleviated. 95% of the doctors agreed that PCa is underdiagnosed at the early stage; hence, the importance of intelligent computational systems for early detection cannot be overestimated.

### 3.2. *PCa data preprocessing environment*

#### 3.2.1. *PCa dataset description*

The dataset was obtained from the Kaggle Healthcare repository (Kaggle, 2022) as comma-separated values (.csv) files. The database contains pretreatment information of patients with PCa; three categories of data merged into one: general background information, serum information, and the SELDI lipid profile. The dataset contains 10,000 medically examined records of patients who are likely to have PCa; it is a labelled dataset consisting of 250 features and one (1) class of features. The background information of patients included age, BMI, race, family history, trouble urinating, urine stream force, blood in semen, bone pain, and erectile dysfunction. The dataset was preprocessed to eliminate class imbalance, incompleteness, noise, and other inconsistencies. The preprocessing included data cleaning, resampling, discretization, and normalization in the following order:

(i) Data cleaning was performed to fill in missing values and discard sparsely distributed records and columns. Data cleaning helped remove noisy data from the dataset, such as missing, inconsistent, and incomplete values.

(ii) Resampling was used to resolve the over- and under-sampling that occurred when dealing with the class imbalance. A class imbalance occurs if class features are unequally represented or distributed. In this study, the target class features considered were PCa and non-PCa. This can be achieved by upsampling the minority class or downsampling the majority class. Both approaches were employed in this study. Before the elimination of the class imbalance, the imbalanced class data were transformed by changing the value from 2 to 1 and 1 to 0, and thereafter read from the dataset to confirm whether the class was imbalanced. Table 2 presents the steps followed to deal with class imbalance using the upsampling and downsampling approaches. When a dataset is used in machine learning, class imbalance makes the results unreliable with poor quality.

(iii) The dataset was discretized by exchanging nominal values with numerical equivalence. This helps reduce the data size and moderate the number of possible differences in each PCa feature. This process fills in the missing values. It also expedites and simplifies the machine learning task.

(iv) Minimum-maximum normalization ensures that the PCa dataset feature is not overwhelmed by other features in terms of distance measure. In this process, the values are adjusted to a range that is usually between 0 and 1. In this study, the minimum-maximum (min-max) normalization method is presented in Eq. (1).

$$f(x) = \frac{x - min\ (x)}{max(x) - min\ (x)} \tag{1}$$

where *min* and *max* are the respective minimum and maximum values of the variable (*feature*) *x,* giving its range. This is known as feature scaling, where the values of the numeric range of a dataset are reduced to a scale between 0 and 1.

The final output of preprocessing is clean, noise-free, consistent, and normalized. The preprocessed dataset contains 9,897 × 236 features and one (1) target class. However, these dimensions are too large for the model to work on. Therefore, the preprocessed dataset needs to be pruned to only the relevant features. This is achieved by ranking the features to predict the target class. This feature-selection and extraction process is known as feature engineering.

#### 3.2.2. *Feature selection and extraction*

Feature selection and extraction are two different techniques that have similar objectives of reducing the dimensions of a dataset. Dimensionality reduction is an important step, particularly for data with many features. Dimensionality reduction is the reduction in random variables under consideration to obtain a set of principal variables (Roweis and Saul, 2000; Tenenbaum et al., 2000).

After cleaning, resampling, discretization, and normalization of the PCa clinical dataset, relevant features were identified using the SelectKBest and chi-squared methods. These were used to reduce the dimensions of our feature sets to relevant features. In addition, to avoid losing data, feature extraction was performed using principal component analysis (PCA). The output was a dataset with reduced dimensions and the target class was preserved. Feature selection ranked our PCa dataset and selected the 12 most relevant features. The advantage of extraction over selection is that the feature sets are utilized more efficiently by feature extraction than by feature selection, which only ranks the features and requires the researcher to select them discretionarily.

*3.2.2.1. SelectKBest technique.* Fig. 2 shows the feature extraction process performed on the PCa dataset using the SelectKBest algorithm. The feature-selection approach used was univariate selection, which uses the SelectKBest class in the Python Scikit-learn library. The SelectKBest class helped score the PCa features using a classifier function, and then removed all but the *k* highest-scoring features. This removed each feature recursively and re-ranked them based on the new features, so that there was a new set $\nabla$ containing $s \in \eta$ elements, so that the algorithm re-ran on the remaining feature sets *k*–1 times; in the end, the new feature sets were qualified for the machine learning model. The analysis of variance *F*-value method was used because the dataset contained both numeric and categorical values. The scores for each attribute were displayed, and the attributes with the highest scores were chosen based on the specified value of *k*.

**Table 2**
Dataset class balancing procedure.

| Class imbalance | Observation |
|---|---|
| **Class** | |
| 1 | 6,435 |
| 0 | 3,565 |
| **New class of upsampling minority class** | |
| 1 | 6,435 |
| 0 | 6,435 |
| **New class of downsampling majority class** | |
| 1 | 3,565 |
| 0 | 3,565 |

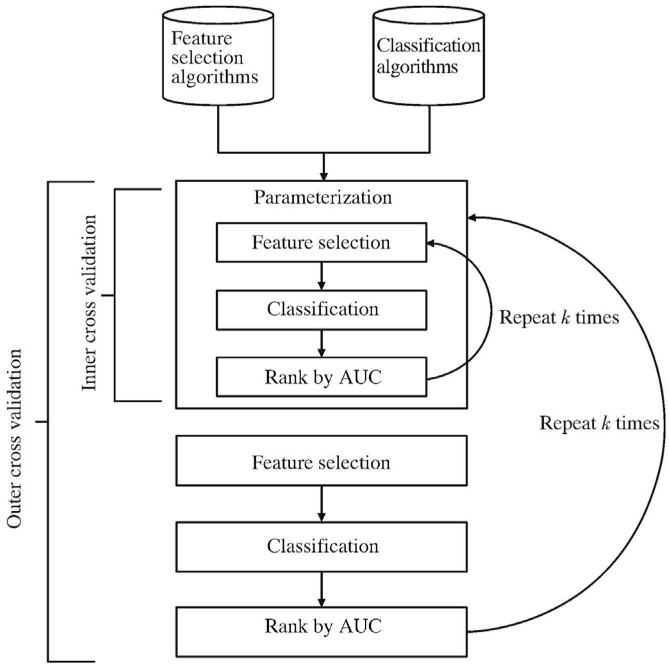

**Fig. 2.** SelectKBest algorithm. AUC: area under the curve.

*3.2.2.2. Chi-squared feature-selection technique.* The chi-square test helped select the best features by testing the relationship between the features and the class. Given each feature set and the corresponding target class, the observed count $O$ and expected count $E$ can be obtained. Chi-square measures how the expected count $E$ and observed count $O$ deviate from each other. This process is denoted by Eq. (2):

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

The $X^2$ of the PCa feature sets $f$ is given by Eq. (3).

$$X^2(f) = \sum_{i=1}^{m} \sum\nolimits_{j=1}^{k} \left( \frac{\left(A_{i,j} - \frac{R_i \times c_j}{N}\right)^2}{\frac{R_i \times c_j}{N}} \right) \tag{3}$$

where:
$m = 250$, number of attributes in the preprocessed PCa database;
$k = 2$, number of classes in the database (six continents);
$N = 10{,}000$, number of samples in the database;
$R_i$, number of patterns in the $i^{\text{th}}$ attribute;
$c_j$, number of patterns in the $j^{\text{th}}$ class;
$A_{i,j}$, number of patterns in the $i^{\text{th}}$ interval and $j^{\text{th}}$ class.
The equation resulting from the above is provided as Eq. (4).

$$X^2(f) = \sum_{i=1}^{251} \sum\nolimits_{j=1}^{2} \left( \frac{\left(A_{i,j} - \frac{R_i \times c_j}{10000}\right)^2}{\frac{R_i \times c_j}{10000}} \right) \tag{4}$$

The attribute with a larger $X^2$ is selected for the classification model. The output of this process is all PCa feature sets with ranking values. The higher the $X^2(f)$ value, the more relevant the feature in the feature set.

*3.2.2.3. PCA.* PCA is a multivariate analysis method based on real eigenvectors. It uses an orthogonal transformation to transform a set of observations that have possibly correlated features. They are transformed into a set of values that have linearly uncorrelated variables, which are called principal components. The transformation is that the first principal component has the highest possible variance, and each succeeding component, in turn, has the highest possible variance under the constraint that it is orthogonal to the preceding components. The higher the PCA value, the more relevant the feature.

*3.2.2.4. Data splitting.* The dataset was partitioned into the training and testing datasets in the ratio of 70:30 for cross-validation. The ratio was experimentally selected because we achieved the optimum prediction accuracy at this ratio. This implies that the training set contained 7,000 instances, and the testing set 3,000 instances. We achieved optimum prediction accuracy based on this ratio (70:30) because it outperformed the prediction accuracy of other ratios, such as 90:10, 80:20, and 67:33, in the experiments we carried out in this study.

### *3.3. Conceptual design of proposed SVM-PCa-EDD*

The SVM model is designed as follows:

(i) Given the training dataset $\theta$ containing feature vectors $x_i$ and their respective class labels $y_i$ in the form of $\theta = [(x_i, \ldots, x_n)] \mid x_i \, \varepsilon \, m^n$, where $m$ is the dimension of the feature vector and $n$ is the number of instances in the dataset;
(ii) Optimal margin classification is achieved by finding a hyperplane in $m$-dimensional space;
(iii) The linear classifier is based on $f(x) = \sum_i w_i \times \ x_i + b$,

where the vector $w_i$ is the weight vector, $b$ is the hyperplane bias, and $f(x) = 0$ is called the hyperplane.

SVM is a supervised machine learning algorithm whose major function is to find the maximum-margin hyperplane for the binary classification of unknown data points from the known classified data point. Regardless of the size of the training set in the domain, new input sets can be predicted faster than in other predictive models. When provided with an unknown PCa tuple without its associated output class, SVM model searches the pattern space for the $K$ training tuples closest to the unknown tuple. The SVM uses classical statistical learning theory and achieves good generalization of new data with a readily interpretable model. The closest points are referred to as support vectors because they support the location of the separating hyperplanes. This means that moving the nonsupport vectors will not change the hyperplanes, and vice versa.

For our PCa dataset, given the training dataset $\theta$ containing feature vectors $x_i$ and their respective class labels, $y_i$, in the form of $\theta = [(x_1, x_2, \ldots, x_n)] \mid x_i \, \varepsilon \, m^n$ where $m$ is the dimension of the feature vector and $n$ is the number of instances in the dataset. In SVM technique, classification is achieved by identifying a hyperplane in an $m$-dimensional space. The linear classifier is based on Eq. (5):

$$f(x) = \sum_i w_i \times \ x_i + b \tag{5}$$

where the vector $w_i$ is the vector of hyperplane coefficients, $b$ is the bias, and with $f(x) = 0$, $x$ is on the hyperplane.

### *3.4. Conceptual diagram of the proposed SVM-PCa-EDD*

Fig. 3 presents a conceptual diagram of SVM-PCa-EDD. The steps involved in model development were data identification and collection, data preprocessing, model development, and model validation. The dataset used was obtained from the Kaggle Healthcare repository and contained various records of the predictive attributes of PCa. Two filter-based feature selections were used to select the best features based on relevance and one feature extraction was used to combine the features before passing the data to the model. Feature selection and extraction were performed on the data to enhance the predictive power of the model and avoid overfitting of the data. The dataset was partitioned into training and testing sets using predefined ratios. The “new tuples” are the data of patients to check their PCa statuses. The data were input into the “validated model” to predict the PCa status per patient. The training set was input into the supervised machine learning algorithm, and the results of the performance were validated based on the predictions made by our models.

### 3.5. *Justification for using SVM*

Deep learning models, such as CNNs, MLPs, RBFNs, RRBFs, RNNs, LSTMs, RBMs, GANs, and DBNs, were not used because our data were not related to cancer image data, image recognition, pattern recognition and prediction, time series and time prediction, visualization, and exploratory data analysis of high-dimensional datasets. The dataset we considered is on early symptoms of PCa because our research focuses on the early diagnosis of PCa before it develops to the formation of cancer lumps. Therefore, our interest is in the dataset of symptoms exhibited by a patient before the formation of PCa lumps. Examples of the symptoms are problems with urination, such as slow or weak stream, frequent urination, especially throughout the night, urinary incontinence or even loss of bladder control, blood in the urine, difficulty getting an erection or having painful erections, and pain in the hips, lower back, and chest.

The Arcene dataset used was obtained from an online repository (Kaggle) as.csv files. It contains pretreatment information of patients with PCa. It contains three categories of data merged in one: general background information, serum information, and SELDI lipid profile of patients. The background information of the patients included age, BMI, race, family history, trouble urinating, urine stream force, blood in semen, bone pain, and erectile dysfunction. Serum information was obtained from the blood samples of potential patients with PCa.

The SVM model was adjusted suitable for the dataset based on the above symptoms. We adopted SVM because it enabled us perform classification and outlier detection. It was used in preprocessing the PCa dataset to mitigate class imbalance and reduce data dimensionality.

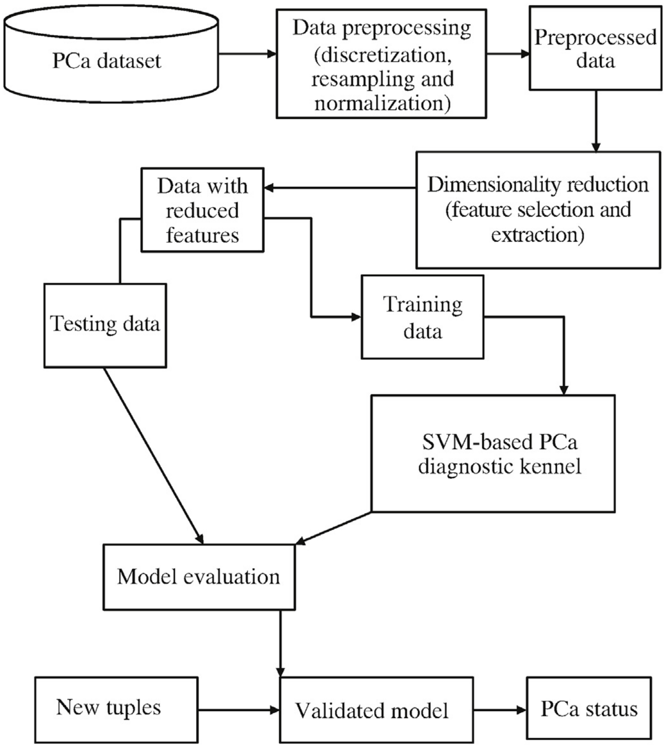
PCa dataset
Data preprocessing (discretization, resampling and normalization)
Preprocessed data
Dimensionality reduction (feature selection and extraction)
Data with reduced features
Testing data
Training data
SVM-based PCa diagnostic kennel
Model evaluation
New tuples
Validated model
PCa status

**Fig. 3.** Conceptual diagram of the proposed support vector machine for early differential diagnosis of Prostate cancer (SVM-PCa-EDD).

### 3.6. *Confusion matrix for performance evaluation of SVM-PCa-EDD*

In this study, the accuracy, sensitivity, specificity, and ROC were used to evaluate the performance of the proposed SVM-PCa-EDD (Table 3). This was obtained from the confusion matrix computed from the prediction output of the training set.

The predicted output was juxtaposed with the actual output. The true positive (TP) is the number of instances predicted to be PCa positive that are actually PCa positive. The false positive (FP) is the number of instances predicted to be PCa positive but actually PCa negative. The false negative (FN) is the number of instances predicted to be PCa negative but actually PCa positive. The true negative (TN) is the number of instances predicted to be PCa negative that are actually PCa negative. Based on the above, the accuracy, specificity, sensitivity, and ROC were computed using Eqs. (6)–(8).

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{d + a}{d + a + b + c} \tag{6}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{d}{d + c} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{a}{a + b} \tag{8}$$

Sensitivity (“positivity in PCa disease”) refers to the proportion of subjects who had the target condition and gave positive test results. Specificity (“negativity in PCa disease”) is the proportion of subjects without the target condition who gave negative test results. All these parameters from the above equations constitute the confidence level of the developed model.

## 4. Results

In this section, the results of the class imbalance and SVM-PCa-EDD implementations are presented and discussed. The experiment was carried out to determine the accuracy of class imbalance and balance class datasets, using logistic regression (LR) with upsampling of the minority class and downsampling of the majority class, and the results were compared. In addition, a comparison of SVM-PCa-EDD with random forest and LR was performed to experimentally validate and test the accuracy of SVM-PCa-EDD. The SVM-PCa-EDD model was implemented using Python. Data preprocessing was performed using data cleaning, normalization, resampling, and discretization. Data cleaning was performed by detecting and correcting (or removing) distorted or inaccurate records from the PCa dataset, as well as identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and replacing, modifying, and deleting dirty or coarse data as necessary, thereby reducing noise. Data transformation was performed before balancing the class using a resampling approach with a random state. The accuracy of class imbalance was 64.4%, the accuracy of upsampling of the minority class was 57.2%, and the accuracy of downsampling of the majority class was 56.5%. The AUC-ROC of the LR model trained on the downsampled dataset was 58.4% while the AUC-ROC of the LR model trained on the class imbalanced dataset was 54.3%. Data were also normalized by

**Table 3**
Confusion matrix template for differential diagnosis of PCa.

| Actual | Target | | | |
|---|---|---|---|---|
| | Prostate | Non-prostate | Predicted value | |
| Prostate (true) | d (TP) | c (FN) | PPV | $\frac{d}{d+b}$ |
| Non-prostate (false) | b (FP) | a (TN) | NPV | $\frac{a}{a+c}$ |
| | Sensitivity | Specificity | Accuracy $= \frac{(d+a)}{(d+a+b+c)}$ | |

Note: positive predicted value (PPV); negative predicted value (NPV).

scaling them to a range of 0–1 using the minimum-maximum algorithm. We discretized the nominal values of the dataset. The result was 10,000 rows × 176 columns.

Chi-square and SelectKBest were used for dimensionality reduction. These two approaches were used for feature selection and extraction. In feature selection, the PCa dataset features were ranked, and the best features were selected for the development of SVM-PCa-EDD. The relevant attributes based on these techniques are features that are regarded as direct predictors of our PCa dataset. The dataset was split in a 70:30 ratio for the training and testing sets.

### *4.1. Results of SVM model on SelectKBest-reduced PCa dataset*

The features from the SelectKBest feature ranking were input into SVM-PCa-EDD using similar parameter tuning. Table 4 presents the performance results of the SelectKBest-reduced dataset.

### *4.2. Results of SVM-PCa-EDD on chi-squared ranking*

The chi-squared output was input into SVM-PCa-EDD, and the performance report is presented in Table 5.

### *4.3. Results of SVM model on extracted features from SelectKBest and chi-squared ranking*

The intersections from the feature-ranking techniques were also taken and passed into the SVM-PCa-EDD prediction engine. The performance is reported in Table 6.

Fig. 4 shows the average error rates of the training model. The error rate of SVM-PCa-EDD helps measure how frequently the algorithm correctly classifies a data point. This is the ratio of correctly predicted data points to all data points. The error rate was reduced with each iteration in the model validation process. In this study, the error rate is the lowest possible error rate for SVM-PCa-EDD, which is often referred to as an irreducible error. The average error rate decreased over time.

### *4.4. Experimental validation of SVM-PCa-EDD with random forest and LR*

A comparison of SVM-PCa-EDD with random forest and LR was performed to experimentally validate the robustness of SVM-PCa-EDD. The comparison results are presented in Tables 7 and 8, and the ROC graph is presented in Fig. 5. The ROCs of LR, SVM-PCa-EDD and random forest were 0.983, 0.986, and 0.985, respectively (Table 8).

**Table 4**
Prediction performance of support vector machine (SVM) engine on SelectKBest-reduced dataset.

| Actual | Target | | | |
|---|---|---|---|---|
| | Prostate | Non-prostate | Predicted value | |
| Prostate | 1,162 | 133 | PPV | 0.94 |
| Non-prostate | 78 | 627 | NPV | 0.83 |
| | Sensitivity | Specificity | Accuracy = 0.90 | |
| | 0.91 | 0.89 | | |

Note: The accuracy was 0.90, with a positive predicted value (PPV) of 0.94, negative predicted value (NPV) of 0.83, specificity of 0.89, and sensitivity of 0.91.

**Table 5**
Prediction performance of SVM-PCa-EDD on chi-squared-reduced dataset.

| Actual | Target | | | |
|---|---|---|---|---|
| | Prostate | Non-prostate | Predicted value | |
| Prostate | 1,173 | 122 | PPV | 0.99 |
| Non-prostate | 75 | 630 | NPV | 0.84 |
| | Sensitivity | Specificity | Accuracy = 0.90 | |
| | 0.91 | 0.89 | | |

Note: The accuracy was 0.90, with a positive predicted value (PPV) of 0.99, negative predicted value (NPV) of 0.84, specificity of 0.89, and sensitivity of 0.91.

**Table 6**
Prediction performance of support vector machine (SVM) engine on intersection (hybrid)-reduced dataset.

| Actual | Target | | | |
|---|---|---|---|---|
| | Prostate | Non-prostate | Predicted value | |
| Prostate | 1,280 | 188 | PPV | 0.96 |
| Non-prostate | 55 | 541 | NPV | 0.74 |
| | Sensitivity | Specificity | Accuracy = 0.88 | |
| | 0.87 | 0.92 | | |

Note: The accuracy was 0.88, with a positive predicted value (PPV) of 0.96, negative predicted value (NPV) of 0.74, specificity of 0.92, and sensitivity of 0.87.

## 5. Discussion

PCa can be diagnosed before it becomes full-blown by analyzing common risk factors using the SVM-PCa-EDD model. The model used risk factors such as age, BMI, urine passage, family history, and bone pain, particularly in the pelvic region. An average accuracy of 90% using the PCA dataset showed that SVM-PCa-EDD is a promising PCa diagnostic model. The results showed that the positive class was better predicted than the negative class. The error rate also demonstrated that the error margin reduces with increasing number of iterations. This shows that the SVM-PCa-EDD can appropriately approximate and predict outcomes, and, therefore, it is reliable. The result also agrees with those of the existing literature (Akinnuwesi et al., 2020; Sand, 2015; Uzoka et al., 2016), which suggests that diseases can be differentially diagnosed at early stages.

The proposed SVM-PCa-EDD helps solve problems attributed to clinical diagnostic methods (i.e., prostate imaging, biopsy, DRE, and PSA), which are used for the detection of PCa. Clinical methods are effective for the detection of PCa; however, they are used at the full-blown stage of PCa. PCa is not detectable by imaging or discovered in bioptic extracts in its early stages when PCa has not yet developed. The

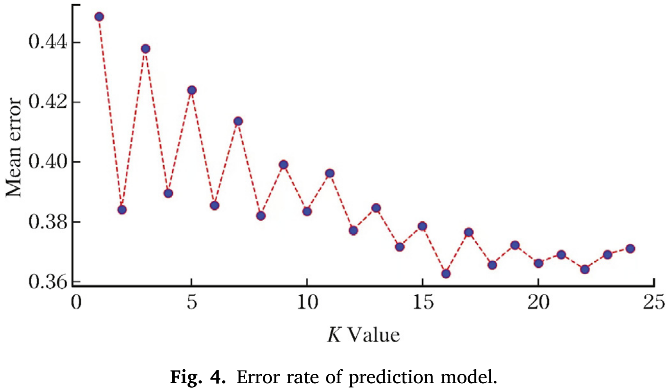

**Fig. 4.** Error rate of prediction model.

**Table 7**
Side-by-side comparison of support vector machine for early differential diagnosis of PCa (SVM-PCa-EDD) with random forest and logistic regression (LR).

| Model | Accuracy | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|
| SVM-PCa-EDD on chi-squared | 0.90 | 0.91 | 0.89 | 0.94 | 0.83 |
| SVM engine on SelectKBest | 0.90 | 0.91 | 0.89 | 0.99 | 0.84 |
| SVM engine on hybrid | 0.88 | 0.87 | 0.92 | 0.96 | 0.74 |

Note: positive predicted value (PPV); negative predicted value (NPV).

**Table 8**
Side-by-side comparison of support vector machine for early differential diagnosis of PCa (SVM-PCa-EDD) with random forest and logistic regression (LR).

| Model | Receiver operating characteristic (ROC) |
|---|---|
| LR | 0.983 |
| SVM-PCa-EDD | 0.986 |
| Random forest | 0.985 |

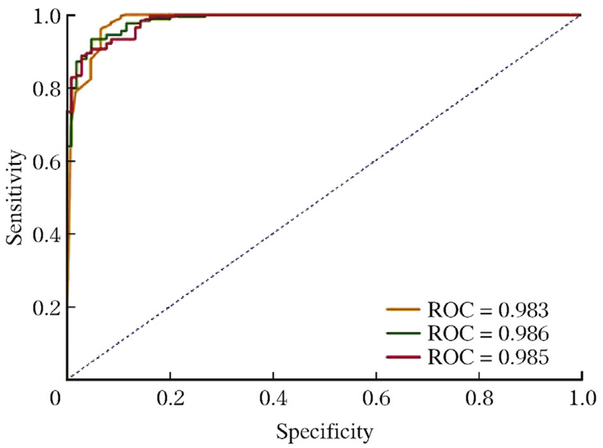

**Fig. 5.** Receiver operating characteristic (ROC) curve of support vector machine for early differential diagnosis of PCa (SVM-PCa-EDD) compared with random forest and logistic regression (LR).

same applies to the use of DRE. Although PSA tests support early PCa detection, the results are not always accurate; therefore, some doctors argue against its use (National Cancer Institute, 2018). Moreover, none of these methods could be used to detect the possibility of PCa in a patient at a very early stage (i.e., when PCa lumps are not yet developed) when the patient exhibits PCa-related symptoms that could be mistaken for some other diseases. The proposed SVM-PCa-EDD could aid in the early detection of PCa, and exhibit 90% accuracy, 80% sensitivity, and 80% specificity.

We also deduced from the reviewed literature that studies conducted on PCa diagnosis relied more on biopsy-based methods than on prostate imaging methods, and most of the studies that relied on biopsy-based methods were restricted to manual diagnosis, although a few recommended ANNs as a computational technique. Moreover, few studies have adopted the imaging/ultrasound diagnosis method, while three different computational methods have been adopted and suggested (i.e., ANNs, fuzzy-based approaches, and cone beam computed tomography). None of the studies in the literature focused on early differential diagnosis of PCa by considering the confusable characteristics of the disease compared to our proposed SVM-PCa-EDD.

In addition, all doctors that we interacted with agreed that an artificial intelligence-based model (such as SVM-PCa-EDD) for the early diagnosis of PCa can be useful. It can serve as a software tool to complement efforts in PCa diagnosis. Adoption, implementation, and utilization of SVM-PCa-EDD in hospitals will help minimize the problem of early-stage PCa underdiagnoses, and thus support doctors in improving their services. The implementation of SVM-PCa-EDD is lightweight and can easily be on any desktop computer or mobile device using Microsoft Windows or Android as the operating system.

## 6. Conclusions

PCa is a chronic and noncommunicable disease that remains a significant global public health problem. Symptoms of PCa are often confused with those of other diseases owing to their similarities, which ultimately leads to late diagnoses and, in some circumstances, underdiagnoses at the early stages. Therefore, in this study, we applied SVM-PCa-EDD, an SVM model, to classify persons with and without PCa using non-image data. A dataset from the Kaggle Healthcare repository was used to develop and validate the model's classification. The SVM-PCa-EDD model was used to pre-process the PCa dataset, dealing with class imbalance and dimensionality reduction. It achieved high performance, with 90% accuracy, 80% sensitivity, and 80% specificity. The validation of SVM-PCa-EDD with random forest and LR shows that SVM-PCa-EDD performed better in early differential diagnosis of PCa.

The results of the validation experiment show that despite the similarities between the symptoms of PCa and those of other diseases that have similar, confusable symptoms, the proposed model (i.e., SVM-PCa-EDD) can, with 90% or higher confidence, predict the likelihood of PCa in patients exhibiting the identified symptoms. It also shows that conducting two or more feature-selection techniques independently and harmonizing their results may further improve the convergence of the SVM-PCa-EDD, thereby improving its accuracy.

The proposed SVM-PCa-EDD method can assist medical experts in early PCa diagnosis, aid targeted treatment, save time, cost, and effort, and minimize PCa-related mortalities. The application of computational systems for early PCa diagnosis is an emerging trend in the fields of medical research, as it helps in computer-aided diagnosis and computer-guided therapy. Computational differential diagnosis can be used to complement conventional clinical methods, especially in resource-constrained healthcare settings.

## Declaration of competing interest

The authors declare that there are no conflicts of interest.

## Acknowledgments

We appreciate the support of the managements of Lagos University Teaching Hospital (LUTH), Nigeria, and Lagos State University Teaching Hospital (LASUTH), Lagos, Nigeria, as well as Dr. Adebisi Ogunjimi, a urological surgeon and consultant in LUTH; Dr. Adegbenga Ajepe, a consultant obstetrician and gynecologist at Parklande Specialist Hospital, Lagos, Nigeria, and Prof. Oluwarotimi Ireti Akinola, an obstetrician and gynecologist at LASUTH. We express our gratitude for supporting us in the process of knowledge acquisition for prostate cancer diagnosis and treatment. We appreciate their assistance in obtaining all the data required for the research.

## References

Adams, J., 1853. The case of scirrhous of the prostate gland with corresponding affliction of the lymphatic glands in the lumbar region and in the pelvis. Lancet 1 (1), 393.

Akinnuwesi, B.A., Adegbite, B.A., Adelowo, F., et al., 2020. Decision support system for diagnosing rheumatic-musculoskeletal disease using fuzzy cognitive map technique. Inform. Med. Unlocked 18 (1), 1–19.

Aldoj, N., Lukas, S., Dewey, M., et al., 2020. Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network. Eur. Radiol. 30 (2), 1243–1253.

Alexander, D.D., Mink, P.J., Cushing, C.A., et al., 2010. A review and meta-analysis of prospective studies of red and processed meat intake and prostate cancer. Nutr. J. 9 (1), 9–50.

Alkadi, R., Taher, F., El-Baz, A., et al., 2019. A deep learning-based approach for the detection and localization of prostate cancer in T2 magnetic resonance images. J. Digit. Imag. 32 (5), 793–807.

Alkhateeb, A., Atikukke, G., Rueda, L., 2020. Machine learning methods for prostate cancer diagnosis. J. J. Cancer 1 (3), 70–75.

American Cancer Society, 2022. American Cancer Society, Inc. Available at: https://www.cancer.org/.

Antonelli, M., Johnston, E.W., Dikaios, N., et al., 2019. Machine learning classifiers can predict Gleason pattern 4 prostate cancer with greater accuracy than experienced radiologists. Eur. Radiol. 29 (9), 4754–4764.

Barentsz, J.O., Richenberg, J., Clements, R., et al., 2012. ESUR prostate MR guidelines 2012. Eur. Radiol. 22 (4), 746–757.

Bernatz, S., Ackermann, J., Mandel, P., et al., 2020. Comparison of machine learning algorithms to predict clinically significant prostate cancer of the peripheral zone with multiparametric MRI using clinical assessment categories and radiomic features. Eur. Radiol. 30 (12), 6757–6769.

Bonekamp, D., Jacobs, M.A., El-Khouli, R., et al., 2011. Advancements in MR imaging of the prostate: from diagnosis to interventions. Radiographics 31 (3), 677–703.

Brabletz, T., Kalluri, R., Nieto, M.A., et al., 2018. EMT in cancer. Nat. Rev. Cancer 18 (2), 128.

Bulten, W., Pinckaers, H., Van Boven, H., et al., 2020. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 21 (2), 233–241.

Bylsma, L.C., Alexander, D.D., 2015. A review and meta-analysis of prospective studies of red and processed meat, meat cooking methods, heme iron, heterocyclic amines and prostate cancer. Nutr. J. 14 (1), 545–552.

Caini, S., Gandini, S., Dudas, M., et al., 2014. Sexually transmitted infections and prostate cancer risk: a systematic review and meta-analysis. Cancer Epidemiol. 38 (4), 329–338.

Cannon, G., Gupta, P., Gomes, F., et al., 2012. Prevention of cancer and non-communicable diseases. Asian Pac. J. Cancer Prev. 13 (4), 3–11.

Cheng, R., Roth, H.R., Lay, N., et al., 2017a. Automatic MR prostate segmentation by deep learning with holistically-nested networks. Medical Imaging 2017. Image Processing. Available at: https://spie.org/Publications/Proceedings/Paper/10.1117/12.2254558.

Cheng, R., Roth, H.R., Lay, N.S., et al., 2017b. Automatic magnetic resonance prostate segmentation by deep learning with holistically nested networks. J. Med. Imag. 4 (4), 041302.

Clark, T., Zhang, J., Baig, S., et al., 2017. Fully automated segmentation of prostate whole gland and transition zone in diffusion-weighted MRI using convolutional neural networks. J. Med. Imag. 4 (4), 1–11.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Cosma, G., Brown, D., Archer, M., et al., 2017. A survey on computational intelligence approaches for predictive modeling in prostate cancer. Expert Syst. Appl. 70 (1), 1–19.

Cunha, G.R., Donjacour, A.A., Cooke, P.S., et al., 1987. The endocrinology and developmental biology of the prostate. Endocr. Rev. 8 (3), 338–362.

De Bono, J.S., Logothetis, C.J., Molina, A., et al., 2011. Abiraterone and increased survival in metastatic prostate cancer. N. Engl. J. Med. 364 (21), 1995–2005.

De Vincentis, G., Monari, F., Baldari, S., et al., 2018. Narrative medicine in metastatic prostate cancer reveals ways to improve patient awareness & quality of care. Future Oncol. 14 (27), 2821–2832.

Denmeade, S.R., Isaacs, J.T., 2002. A history of prostate cancer treatment. Nat. Rev. Cancer 2 (5), 389.

Federman, D.G., Chanko, E.H., 2007. Differential Diagnosis in Internal medicine: from symptom to Diagnosis. JAMA 298 (17), 2070–2075.

Gann, P.H., 2002. Risk factors for prostate cancer. Rev. Urol. 4 (Suppl. 5), S3–S10.

Gerlinger, M., Rowan, A.J., Horswell, S., et al., 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. 366 (10), 883–892.

Goldenberg, S.L., Nir, G., Salcudean, S.E., 2019. A new era: artificial intelligence and machine learning in prostate cancer. Nat. Rev. Urol. 16 (7), 391–403.

Hagiwara, Y., Fujita, H., Oh, S.L., et al., 2018. Computer-aided diagnosis of atrial fibrillation based on ECG signals: a review. Inf. Sci. 467 (Oct.), 99–114.

Houston, K.A., King, J., Li, J., et al., 2018. Trends in prostate cancer incidence rates and prevalence of prostate specific antigen screening by socioeconomic status and regions in the United States, 2004 to 2013. J. Urol. 199 (3), 676–682.

Kaggle, 2022. Prostate cancer. Available at: https://www.kaggle.com/datasets/sajidsaifi/prostate-cancer.

Kakade, S.M., Sridharan, K., Tewari, A., 2009. On the complexity of linear prediction: risk bounds, margin bounds, and regularization. In: Proceedings of the 21st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, pp. 793–800.

Kote-Jarai, Z., Al Olama, A.A., Giles, G.G., et al., 2011. Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. Nat. Genet. 43 (8), 785.

Kott, O., Linsley, D., Amin, A., et al., 2021. Development of a deep learning algorithm for the histopathologic diagnosis and Gleason grading of prostate cancer biopsies: a pilot study. Eur. Urol. Focus 7 (2), 347–351.

Kumagai, H., Zempo-Miyaki, A., Yoshikawa, T., et al., 2015. Lifestyle modification increases serum testosterone level and decrease central blood pressure in overweight and obese men. Endocr. J. 62 (1), 423–430.

Kwak, J.T., Hewitt, S.M., 2017. Lumen-based detection of prostate cancer via convolutional neural networks. Medical Imaging 2017: Digital Pathology. Availlable at: https://doi.org/10.1117/12.2253513.

Li, C., Chen, M., Wan, B., et al., 2018a. A comparative study of Gaussian and non-Gaussian diffusion models for differential diagnosis of prostate cancer with in-bore transrectal MR-guided biopsy as a pathological reference. Acta Radiol. 59 (11), 1395–1402.

Li, W., Li, J., Sarma, K.V., et al., 2018b. Path R-CNN for prostate cancer diagnosis and gleason grading of histological images. IEEE Trans. Med. Imag. 38 (4), 945–954.

Liberman, M.C., Epstein, M.J., Cleveland, S.S., et al., 2016. Toward a differential diagnosis of hidden hearing loss in humans. PLoS One 11 (9), 1–15.

Liu, B., Cheng, J., Guo, D., et al., 2019. Prediction of prostate cancer aggressiveness with a combination of radiomics and machine learning-based analysis of dynamic contrast-enhanced MRI. Clin. Radiol. 74 (11), 896.

Lokhande, A., Bonthu, S., Singhal, N., 2020. Carcino-Net: a Deep learning Framework for automated Gleason grading of prostate biopsies. In: 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, Montreal.

Mann, R.S., 1990. Differential diagnosis and classification of apathy. Am. J. Psychiatr. 147 (1), 22–30.

Marley, A.R., Nan, H., 2016. Epidemiology of colorectal cancer. Int. J. Mol. Epidemiol. Genet. 7 (3), 105.

Moyad, M.A., 2002. Is obesity a risk factor for prostate cancer, and does it even matter? A hypothesis and different perspective. Urology 59 (4), 41–50.

Mumford, C., Jain, L., 2009. Computational intelligence: collaboration: fusion and emergence. Intelligent Systems Reference Library. Springer, Berlin.

Murtola, T.J., Vihervuori, V.J., Lahtela, J., et al., 2018. Fasting blood glucose, glycaemic control and prostate cancer risk in the Finnish Randomized Study of Screening for Prostate Cancer. Br. J. Cancer 118 (9), 1248–1254.

Nair, R.V., Vijaya, R., 2010. Photonic crystal sensors: an overview. Prog. Quant. Electron. 34 (3), 89–134.

Naji, L., Randhawa, H., Sohani, Z., et al., 2018. Digital rectal examination for prostate cancer screening in primary care: a systematic review and meta-analysis. Ann. Fam. Med. 16 (2), 149–154.

Natarajan, S., Marks, L.S., Margolis, D.J., et al., 2011. Clinical application of a 3D ultrasound-guided prostate biopsy system. Urol. Oncol.: seminars and original investigations 29 (3), 334–342.

Ng, K.L.S., Mishra, S.K., 2007. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics 23 (11), 1321–1330.

Nir, G., Karimi, D., Goldenberg, S.L., et al., 2019. Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images. JAMA Netw. Open 2 (3), e190442.

Parikesit, D., Mochtar, C.A., Umbas, R., et al., 2016. The impact of obesity towards prostate diseases. Prostate Int. 4 (1), 1–6.

Pereira, M.M., Calixto, J.D., Sousa, A.C., et al., 2020. Towards the differential diagnosis of prostate cancer by the pre-treatment of human urine using ionic liquids. Sci. Rep. 10 (1), 1–8.

Platz, E.A., Till, C., et al., 2009. Men with low serum cholesterol have a lower risk of high-grade prostate cancer in the placebo arm of the prostate cancer prevention trial. Cancer Epidemiol. Prevent. Biomark. 18 (11), 2807–2813.

National Cancer Institute, 2018. Prostate-specific antigen (PSA) test. Available at: https://www.cancer.gov/types/prostate/psa-fact-sheet.

Rahman, M.T., Chowdhury, A.M., 2016. Prostate cancer. Anwer Khan Mod. Med. Coll. J. 7 (2), 36–44.

Rice, S.B., Nenadic, G., Stapley, B.J., 2005. Mining protein function from text using term-based support vector machines. BMC Bioinf. 6 (1), 1–11.

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Sand, I.K., 2015. Classification, diagnosis, and differential diagnosis of multiple sclerosis. Curr. Opin. Neurol. 28 (3), 193–205.

Schoots, I.G., Roobol, M.J., Nieboer, D., et al., 2015. Magnetic resonance imaging-targeted biopsy may enhance the diagnostic accuracy of significant prostate cancer detection compared to standard transrectal ultrasound-guided biopsy: a systematic review and meta-analysis. Eur. Urol. 68 (3), 438–450.

Siddiqui, M.M., Rais-Bahrami, S., Turkbey, B., et al., 2015. Comparison of MR/ultrasound fusion-guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer. JAMA 313 (4), 390–397.

Silva-Rodríguez, J., Colomer, A., Sales, M.A., et al., 2020. Going deeper through the Gleason scoring scale: an automatic end-to-end system for histology prostate grading and cribriform pattern detection. Comput. Methods Progr. Biomed. 195 (Oct.), 1–44.

Smith, M., Coleman, R., Klotz, L., et al., 2015. Denosumab for the prevention of skeletal complications in metastatic castration-resistant prostate cancer: comparison of skeletal-related events and symptomatic skeletal events. Ann. Oncol. 26 (2), 368–374.

Steinestel, J., Luedeke, M., Arndt, A., et al., 2019. Detecting predictive androgen receptor modifications in circulating prostate cancer cells. Oncotarget 10 (41), 4213.

Sung, H., Ferlay, J., Siegel, R.L., et al., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 71 (3), 209–249.

Taitt, H.E., 2018. Global trends and prostate cancer: a review of incidence, detection, and mortality as influenced by race, ethnicity, and geographic location. Am. J. Men's Health 12 (6), 1807–1823.

Tenenbaum, J.B., De Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Tian, W., Osawa, M., 2015. Prevalent latent adenocarcinoma of the prostate in forensic autopsies. J. Clin. Pathol. Forensic Med. 6 (3), 11–13.

Tian, Z., Liu, L., Fei, B., 2018. Deep convolutional neural network for prostate MR segmentation. Int. J. Comput. Assist. Radiol. Surg. 13 (11), 1687–1696.

To, M.N.N., Vu, D.Q., Turkbey, B., et al., 2018. Deep dense multi-path neural network for prostate segmentation in magnetic resonance imaging. Int. J. Comput. Assist. Radiol. Surg. 13 (11), 1687–1696.

Tse, L.A., Lee, P.M.Y., Ho, W.M., et al., 2017. Bisphenol A and other environmental risk factors for prostate cancer in Hong Kong. Environ. Int. 107 (Oct.), 1–7.

Turner, N., Reis-Filho, J., Russell, A., et al., 2007. BRCA1 dysfunction in sporadic basal-like breast cancer. Oncogene 26 (14), 2126–2132.

Uzoka, F.M.E., Akinnuwesi, B.A., Amoo, T., et al., 2016. A framework for early differential diagnosis of tropical confusable diseases using the fuzzy cognitive map engine. World Academy of Science, Engineering and Technology, Int. J. Comput. 10 (2), 346–353.

Vaidyanathan, V., Naidu, V., Kao, C.H.J., et al., 2017. Environmental factors and risk of aggressive prostate cancer among a population of New Zealand men–a genotypic approach. Mol. Biosyst. 13 (4), 681–698.

WHO, 2020. Cancer, Available at:. https://www.who.int/news-room/fact-sheets/detail/cancer.

Wolk, A., 2017. Potential health hazards of eating red meat. J. Intern. Med. 281 (2), 106–122.

Yaghi, M.D., Kehinde, E., 2015. Oral antibiotics in trans-rectal prostate biopsy and its efficacy to reduce infectious complications: systematic review. Urol. Ann. 7 (4), 417.

Yaxley, J.W., Coughlin, G.D., Chambers, S.K., et al., 2016. Robot-assisted laparoscopic prostatectomy versus open radical retropubic prostatectomy: early outcomes from a randomised controlled phase 3 study. Lancet 388 (10049), 1057–1066.

Zhang, J., Dong, M., Hu, X., et al., 2016. Prostatic adenocarcinoma presenting with metastases to the testis and epididymis: a case report. Oncol. Lett. 11 (1), 792–794.

Zhou, C.K., Check, D.P., Lortet-Tieulent, J., et al., 2016. Prostate cancer incidence in 43 populations worldwide: an analysis of time trends overall and by age group. Int. J. Cancer 138 (6), 1388–1400.