VER
Vascular & Endovascular Review

Prediction of Prostate Cancer by Using Data Mining Techniques with Artificial Neural Network

Mohammad Hassan Osman Abdalraheem[1], *, Alfadil Ahmed Hamdan[2], Mohand Tahar Kechadi[3]

[1]Department of Computer Science, College of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia. Email: mohammedh@jazanu.edu.sa, ORCID: https://orcid.org/0000-0002-0013-0505
[2]Faculty of Computer Science and Information Technology, Sudan University of Science and Technology, Sudan. Email: alfadil.hamdan@sustech.edu; dr.alfadil2012@gmail.com
[3]School of Computer Science, University College Dublin, Ireland. Email: tahar.kechadi@ucd.ie

***Corresponding author**: mohammedh@jazanu.edu.sa

ABSTRACT

The idea is to explore the potential of using machine learning algorithms in predicting prostate cancer. The estimation models of prostate cancer should be made appropriately to increase the chance of prostate cancer survival. Prostate-cancer is a highly prevalent cancerous disease in men. Therefore, recognition of prostate-cancer as early as possible is needed and essential so that practitioners can establish probable outcomes and set effective strategies to treat the patients. The dataset used was the original database with 76,683 records of patients. The data were acquired from the National Cancer Institute (NCI). The number of patients after processing, cleaning, and eliminating duplicate data stands at 36,159. The data for the produced artificial neural network (ANN) was provided through the selection of data attributes, e.g., depending on the patient's age and the levels of PSA. We used a methodology that divided the data into 70% training and 30% test and validation sets, respectively. AUC and ROC curves are applied to evaluate the experimental outcomes of the proposed model. An ROC curve is a standard method of illustrating the performance of a binary classifier in terms of its accurate positive and false positive rates over a threshold range. The AUC provides a point of reference, giving a single number to represent the overall performance. Our in-depth ANN, educated by means of scaled conjugate convergence feed-forward backpropagation, achieved 91.50% in the overall confusion matrix. The best validation is 0.0644 at measure 26, the gradient is 0.0151 at measure 32, and the validation check is six at measure 32. ANN was close in predicting prostate cancer and did not use a biopsy. Nevertheless, we believe that ANN performance is insufficient for application in clinical practice. This paper observed that prostate cancer can be effectively predicted using machine learning methods. Nevertheless, incorporating more variables, including life factors that are suitable to estimate prostate-cancer, could also be a good alternative that would enable the ANN to perform better.

KEYWORDS: Machine Learning; Artificial Neural Networks; Prostate Cancer; Prostate Specific Antigen.

How to Cite: Mohammad Hassan Osman Abdalraheem, Alfadil Ahmed Hamdan, Mohand Tahar Kechadi, (2026) Prediction of Prostate Cancer by Using Data Mining Techniques with Artificial Neural Network, Vascular and Endovascular Review, Vol.9, No.1, 125-134

INTRODUCTION

The most visible cancer in men is prostate cancer, besides lung cancer. It is estimated that prostate cancer would result in almost 174,650 instances and 31,620 deaths in 2019 in the United States. It causes virtually every 1 in 5 new cancer cases in men [1][2]. Early and proper diagnosis would mean a lot in lessening the death rate due to prostate cancer. The most common way of conducting a diagnosis of prostate cancer clinically is to conduct a needle biopsy [3][4]. A radiologist specializing in genitourinary tract should conduct prostate biopsies, which are painful and may be marked with the risks of infection & bleeding. It is, therefore, not an excellent choice for diagnosing patients. Having an increased serum concentration of PSA is used as a common biomarker of needle biopsy to help in the early identification of prostate cancer[5]. A PSA is an antigen related to the prostate gland.

Nevertheless, the PSA screening exposes a large proportion of men to the risk of misdiagnosis and unnecessary treatment [6], which leads to the repetitive biopsy of many patients with cancer in other locations. Machine learning has gained popularity in clinical medicine over the past few years, particularly as a means of creating prediction models [7][8]. To this end, we intend to employ machine learning to forecast prostate cancer using data gathered by the NCI (National Cancer Institute) [9], with PSA as one of the central variables. To enhance the performance of prediction, we utilized machine learning algorithms. Machine learning (ML) is a subject within data science and artificial intelligence [10] that focuses on the development and training of algorithms through which computers can learn and make predictions based on data without explicit programming [11]. To grow a reliable and precise pattern for prostate cancer estimation, the proposed paper has developed an artificial neural network algorithm trained using PSA (prostate-specific antigen) as a biomarker [12][13].

Prostate cancer is among the most widespread and visible forms of cancer in men all over the world, and as far as the United States is concerned, it comes second only to lung cancer in terms of occurrence. In 2019, epidemiological estimations show that about 174,650 cases of “prostate cancer “were detected in the year 2019 and 31,620 men died because of the disease. Prostate cancer is a major health issue affecting a great percentage of the male population because of its high rate. It is important to note

that the statistics show that almost 20 percent (or 1 in 5) of all the fresh cancer cases of men are linked to prostate cancer [14][15]. This prevalence makes it very important to note that early detection and proper diagnosis techniques are of vital importance, as the earlier the disease is detected, the better the treatment results and the less likely to die.

Another method of diagnosing prostate cancer is the traditional gold standard, which is the needle biopsy. It involves the insertion of a fine indicator throughout the rectal border or perineum to remove small bits of tissue tests from various spots of prostate gland to undergo microscopic analysis [16]. The biopsy is usually performed by a skilled radiologist or urologist specializing in genitourinary imaging with the assistance of ultrasound imaging in order to specify the suspicious areas. Although this is the best technique to give conclusive histopathological investigations of malignancy, it is invasive and can be painful to the patient [17][18]. In addition, needle biopsies are associated with the risks of infection, bleeding, pain, and swelling that may cause great discomfort and anxiety, and sometimes may result in the unwillingness of patients to undergo the tests required to be conducted to diagnose [19][20]them.

Considering such constraints, biomarkers are widely used by healthcare practitioners to indicate the possibility of prostate cancer [21] and assess the need to conduct a biopsy or other tests. PSA is the most frequently utilized and accessible biomarker and is a protein secreted by prostate tissue. An increased level of PSA in blood may be a sign of prostate cancer, benign prostatic hyperplasia (BPH) or prostatitis. PSA screening is a non-invasive test that is comparatively cheap and is widely accessible, hence it is the first in a series of diagnostic measures of prostate cancer. PSA testing is, however, not foolproof. Its sensitivity is high, hence most men who have high levels of PSA, but no cancer, are likely to undergo unnecessary biopsies, which expose them to possible complications as well as psychological distress.

Moreover, PSA levels may be affected during different benign disorders, which is why false positives may be provided. On the other hand, there are cases of men with prostate cancer having a normal PSA level, giving false negatives and failure to diagnose prostate cancer. Such ambiguity makes clinical decision-making challenging and may result in overdiagnosis, excessive treatment, or tardy diagnosis, potentially with a negative impact on patient outcomes [22].

The medical profession has increasingly resorted to new technologically advanced computational methods, in particular, machine learning (ML), in the quest to enhance early detection [23]. ML consists of algorithms and statistical models that allow computers to learn patterns in data, predict or categorize data without specific instructions to do so. The method is especially beneficial in learning more complicated (or higher-dimensional) data in the medical field, including genomic, imaging, lab, and clinical variables.

ML has been making huge momentum in the field of clinical medicine over the last few years, and prostate cancer is no exception. Its ability to process large and diverse data enables it to build predictive models that can guide clinicians in improving diagnoses by making them more accurate and personalized. Machine learning algorithms [24] can enhance the accuracy of risk stratification, prognosis prediction, and treatment planning by examining patterns that would otherwise be undetected using conventional statistical techniques.

We hope to capitalize on these benefits in our research by coming up with powerful predictive models of detecting prostate cancer relying on data obtained at the National Cancer Institute (NCI) [4]. We mainly consider the inclusion of PSA levels with other pertinent clinical and demographic factors, which family history, include age, race, prostate volume, and other biomarkers. With such varied characteristics, we will develop a multifaceted model that will enable us to estimate accurately the likelihood of a person having prostate cancer.

In particular, we offer to use sophisticated Artificial Neural Network (ANN) software. ANNs are based on the layout of the individual brain and are connected layers of nodes (or neurons) that have the ability to represent complex, non-linear relationships within data. They are exceptionally well adapted to medical diagnosis because of their capacity to learn complex patterns and several variables react in a non-linear manner. An ANN can be trained by providing it with large volumes of labelled data, i.e. cases of assured prostate cancer diagnosis and control cases and letting the network optimally adjust its internal parameters to allow it to maximize its prediction accuracy.

Through the design and training of such an ANN, we plan to identify any hidden patterns in PSA levels and other clinical markers that will differentiate true positive cases from false positives or negatives. The intended result is the development of a non-invasive, affordable and high-quality screening device capable of aiding clinicians in decision-making, minimizing the total of redundant biopsies, as well as early detecting prostate cancer. With this, we will have a chance of helping to improve patient outcomes, less invasive diagnostic pathways, and resource utilization in the management of prostate cancer.

This paper is organized in the following way. Section 2 describes the associated work on methods for early recognition of prostate cancer that employ machine learning systems. Section 3 explains the problem and suggests a method of solving it. Section 4: investigational outcome and discussions. The final part is the conclusion, which proposes a few areas that may be explored in the future.

RELATED WORKS

Over the years, many methods have been developed regarding the diagnosis and prediction of prostate cancer through screening. Introduced in 1994, ANNs became the first method of prostate cancer diagnosis because they can be defined as a machine learning technique [5]. They employed three input data points (age, level of PSA, arithmetical rectal examination (DRE), and TRUS

results) and achieved an outstanding AUC of 0.87. ANNs have been popular since then for this application. PSA analysis is one of the first detection tests. In another study conducted by authors [16], different cut-offs for PSA level recognition of prostate cancer are used, with a PSA (3.1 ng/ml, which reveals a sensitivity result of 0.32 and a specificity result of almost 0.87. The result of a study by researchers [25] was to foresee the possibility of prostate cancer and to calculate the accuracy of the predictions as well as the ROC operating characteristic by using DL and multilayer ANN(MANN) programs and training the model using the data of 232 patients and testing it using 102 patients. This is because they established that artificial neural networks (ANNs) are able to predict prostate cancer even without biopsy, and their performance was slightly better compared to the linear regression method (LR). But, to act practically in the clinic, the performance of the ANN might be further improved. Roffman et al. (2018 developed an ANN to model the risk exist in prostate cancer depending on the evidence such as BMI, age, diabetes status, emphysema, smoking status, race, asthma, ethnicity, cardiovascular disease, HTN, exercise, and stroke history by means of the National Health Interview Survey (NHIS) adult study on the training and test sets.

In recent decades, there have been major improvements in the procedures of diagnosis and prediction of “prostate cancer” with the help of different screening procedures. Among them, the Artificial Neural Networks (ANNs) became one of the pioneering methods in the year 1994 and served a turning point in the use of machine learning in medical diagnostics. ANNs are computing algorithms that are based on the “biological neural networks”(BNN) of the human brain, which are able to discover complex patterns in large datasets. When used in detecting prostate cancer, their introduction was groundbreaking due to their capability to process several variables at a time and learn to use the information to enhance diagnostic accuracy.

The first ANNs used in the diagnosis of prostate cancer used only three input features, which included the age of the patient, the amount of PSA, a “digital rectal examination” (DRE) and “transrectal ultrasound” (TRUS) results. Although these input variables were simple, the neural network models had eminent performances with an AUC of 0.87 in “Receiver Operating Characteristic” (ROC) analysis. Such a large AUC means that it is very effective at separating benign and malignant cases, which further makes ANNs appropriate in this context. Since then, neural networks have continued to be popular in screening prostate cancer because of their flexibility and capacity to screen multiple sources of data.

One of the oldest and most prevalent tests to detect prostate cancer is PSA analysis. However, the PSA levels cannot be fully used as there is a crossover of the values between healthy and prostate cancer patients. As an example, authors studied the various cutoff thresholds of PSA to enhance the accuracy of detection [26]. They discovered that a sensitivity of about 32 was obtained with a threshold PSA level of 3.1 ng/ml. Sensitivity is the capacity of the test to detect the presence of individuals with the disease correctly; however, in the present case, it signifies that a significant proportion of instances might be missed. The specificity, however, or the capability of correctly identifying the disease-free individuals, was high at approximately 87 percent, which minimizes the false positives.

However, an overview of research done by works [27][28] demonstrated how deep learning (DL) and multilayer ANNs (also known as Multilayer Artificial Neural Networks or MANNs) could be used to predict prostate cancer. They were trained on a dataset of 232 patients and tested on another test dataset of 102 patients. The models were not only used to predict the danger of “prostate cancer”, but also to optimally measure the performance of each model using such measurements as the area of ROC curve (AUC). Their findings revealed that ANNs were able to forecast prostate cancer [29][30] with marginally higher precision than conventional models of linear regression (LR). Notably, these patterns demonstrated a potential for non-invasive diagnosis that has the potential to minimize invasive tests like biopsies.

Regardless of these advances, clinical applicability would still need more advancement in the strength and precision of ANN models. In this regard, researchers [31] created an ANN model that incorporated a larger number of variables related to the risk of prostate cancer. Their model included body mass index (BMI), age, whether they had diabetes, emphysema, whether they were a smoker, their race, whether they belonged to an ethnic group and had cardiovascular disease, high blood pressure (HTN), how often they exercised and whether they had a stroke or not. They used national data from the “National Health Interview Survey” (NHIS), which is a significant-scale countrywide dataset, to train and test their model on different datasets. This multifactorial strategy was intended to have the capacity to embrace the wide range of individual health and demographic aspects that have a role in causing prostate cancer and hence enhance the predictive ability and clinical utility [32].

These efforts are in line with the general objective of our present work. Our goal is to design and enhance machine learning algorithms, including ANNs, to detect the presence or the absence of prostate cancer using medical data easily obtained. More importantly, our models will be designed in such a way that they will not require invasive procedures such as biopsies, which may be painful and risky to patients. Through the use of the current variables of the routine medical check-ups and medical history, we are aiming to offer an effective, non-invasive screening tool that will be able to help practitioners in early diagnosis and a customized approach to management of patients.

Through our work, our aim is to train ML algorithms to identify the existence or absence of “prostate cancer” established on the existing variables within the medical data, which can be determined without the need for biopsies.

METHODOLOGY

3.1 Data Description

In our project, we constructed a machine learning scheme using prostate data obtained from the “National Cancer Institute” (NCI) Prostate Cancer Trial (PLCO)—the initial records comprised 76,683 bank patients in the original database. The dataset embraces respective Excel files containing information on PSA testing, biopsy testing, blood repetitive routine examination, transrectal

ultrasonography, urine routine check, gonadal hormone levels, biochemical examination, and diagnostic details. Each file has several characteristics, like patient ID, the time & date of the examination, various testing indexes for each examination, etc. All files are associated with the patient ID and the date of the exam.

3.2 Data Preprocessing

Preprocessing of the data is an emergent process in the prediction of "prostate cancer" through data mining. It entails the process of cleaning and changing raw data so that it becomes more acceptable and comfortable to analyze. The process involves processing the missing values, the elimination of noise and outliers, normalization / scaling the features (PSA levels and age) and the encoding of categorical variables. Preprocessing promotes the predictive model accuracy and reliability to the extent that data that is processed into algorithms, such as machine learning or data mining tools, is uniform, comprehensive, and properly organized.

Our dataset has been prepared using processes that may include the following steps:

- Remove redundant (duplicate data) using MATLAB software
- For missing values handling, we fill all the variables that have a null value with zero
- Selected some features according to our requirements.

3.3 Machine Learning Method

This work provided a supervised machine learning model, which was an Artificial Neural Network (ANN). A set of input variables such as patient age, dreresult0-3 Digital Rectal Examination results of each screening, psaresult0-5 Prostate-Specific Antigen results of each screening and PSA level recorded from a valid PSA screening in each screening. Table 1 presents the traits of the data.

Table 1: Data characteristics according to the results of pathological analysis

Attributes	Description	Data values
age level	Definite version of age, established from the obtained age variable.	0 is <= 59 1. between 60 & 64 2. between 65 & 69 3. more than equal to 70
DREresult0	Result of the DRE for every screen	C= "Control" 1= "Negative" 2= "Abnormal, suspicious" 3= "Abnormal, non-suspicious" 4= "Inadequate screen" 8= "Not done, expected" 9= "Not done, not expected"
DRE result1		
DREresult2		
DREresult3		
PSAresult0	PSA result screen wise	C="Control" 1="Negative" 2="Abnormal, suspicious" 4="Inadequate screen" 8="Not done, expected" 9="Not done, not expected"
PSAresult1		
PSAresult2		
PSAresult3		
PSAresult4		
PSAresult5		
PSAlevel0	PSA readings obtained from an accurate PSA monitor for every monitor	Numeric 6.1 .C="Control". N="Not Applicable"
PSAlevel1		
PSAlevel2		
PSAlevel3		
PSAlevel4		
PSAlevel5		

The proposed model has 17 input attributes and one predictor attribute as output. The number of patients is 36,159 instances after preprocessing, handling missing values, and eliminating redundancy. We have divided the data into two parts: 70% of the data is used to train the model, and the remaining 30% used to test and validate the model. We developed an ANN and analyzed it with MATLAB Neural Network Toolbox software.

3.3.1 Artificial Neural Network (ANN)

The ANNs are computing models that is encouraged by the shape and behavior of biological neural networks. They are set to identify patterns, learn through information and make forecasts or decisions. ANNs are regarded as learning-based algorithms in the sense that they adapt their internal parameters, namely, the weights of connections between nodes, according to the input information that they receive, which allows them to become more efficient with time. This learning through adaptation has been compared to biological neural networks, which include genetic neural networks that learn and adapt like the human brain.

The processing unit in ANN is made up of neurons or connexons that are arranged in layers. One of the most used structures of an ANN is the input layer, the hidden layer/layers, and an output-layer. The connections between neurons have a weight which defines the strength and the impact of the signal going through that connection. These connections carry the input data during the operation; each neuron adds a biased sum of its inputs. In most cases, the sum is managed by an activation work to give the neuron its output. The processed signals are further transmitted over the network up until they link to the yield layer, which

produces the final output.
The essential feature of neural networks is that they are flexible in terms of input data. They can understand intricate, non-linear relations, and that is why they are very applicable in the tasks of classification, regression, pattern recognition, and prediction. There are also two key steps in the development and deployment of an ANN: training and testing.

a. Training Phase:
The neural network learns to give out desired outputs when the required inputs are given during training. It does that through the manipulation of its weights to reduce the gap between its forecasted output and the real or anticipated output, also referred to as the label. Learning is achieved by exposing the network to a high number of labelled data in order to enable the network to iteratively adjust the weights to improve the mapping of the inputs to the outputs. This is what is commonly termed as supervised learning.

In order to optimize the weights, neural networks use algorithms like the gradient descent, which calculate the gradient (i.e. the slope) of a loss function with respect to each weight. The idea is to move the weights in a small step towards the direction of minimizing the total error. This entails computing the error signals through a procedure termed error backpropagation and the weight revision. The training data is run through the network repeatedly (epochs) until the performance of the network reaches a satisfactory level or a set of weights which is optimal.

It is prominent to observe that the neural network should be trained over a period of time before it can be applied to real-world issues, i.e. classifying images, diagnosing problems, or predicting future trends, as well-trained weights allow proper and dependable performance.

b. Testing Phase:
After the training, the neural network passes into the testing/inference phase. In this case, the trained model is given new and unknown data to test its predictive performance. The network takes the input data, passes the signals forward across the layers (feedforward) and produces an output prediction. In the testing phase, the network is tested on its capacity to generalize on new data. In case the productivity of the network is close to the anticipated output within a given tolerance threshold, the training can be terminated or optimized further. The testing stage assists in establishing whether the network is overfitting or underfitting the training data.

c. Popular Algorithms:
Backpropagation (BP) is the most popular algorithm to use in training neural networks. This algorithm is effective in calculating the descent of the deficit function with regard to each weight, and this facilitates effective updates on the weights. BP consists of two main steps:

d. Feedforward Pass:
The network takes the input data, layer by layer, until an output is generated. This step entails the summation of the weighted input of the individual neurons, and the activation functions (tanh, ReLU, or sigmoid) are employed to provide non-linearity.

e. Backpropagation Step:
The network measures its performance against a predetermined performance target and determines the error. The fault is afterwards backpropagated through the network layer by layer, with the weights being updated to reduce the overall error. The error signals are used to adjust the weights in the gradient descent method that attempts to locate the minimum value of the error surface by moving in the inverse route of the gradient.

f. The Gradient Descent Algorithm:
A gradient descent is used to update the weights by moving in the direction that iteratively minimizes the loss. Practically, a learning rate scales the magnitude of every step, aiding in balancing both the speed of convergence and stability. This cycle is repeated until the error, commonly defined as the Mean Squared Error (MSE), does not decrease or start to grow, or the overfitting is visible.

The ANN is a learning-based algorithm whose attributes are based on biological ones, as in genetic neural networks and is composed of a collection of interconnected input & output units, with every interconnection having its characteristic weight. The connected neurons send messages to one another, and their numerical weight can be adjusted based on the experience that the neural network has. Hence, neural networks are flexible to input and can learn [9]. Development of an artificial neural network (ANN) takes place in two stages: training and testing. The training stage involves the network learning to produce the expected output in response to an input in the form of data. In the learning aspect, neural networks are trained to learn by optimizing the weights to estimate the label of the input records, which is a class. Neural networks should be trained extensively before their application to a specific problem, such as classification or prediction [10].

Conversely, at the testing tier, we test the network's capabilities to halt or save training and can use it to predict an output. The network ceases its training when the tested fault is within the desired tolerance value. The algorithm that is most popular is the backpropagation (BP) [11]; its area of application is the widest. The algorithm consists of two steps, namely the feed-forward and backpropagation steps. The feedforward utilizes information to create an output established on the input layer to the output layer. In backpropagation, the system compares the actual output obtained through feed-forward and the expected output and then computes the mistake of the “output layer”. The error thereby generated is passed back to regenerate the connections on the input

layer. The algorithm that is applied to BP exercise is the gradient descent algorithm. The BP algorithm aims at ensuring that the network performs better by minimizing the overall error as much as possible by carrying out the weight updates that are carried along the gradient. Training is paused when the "mean squared errors" (MSE) no longer improve, but the increase [12].

3.3.2 Application of artificial neural network to patient's dataset

The reason to apply ANN in the given work is to identify whether "prostate cancer" exists based on the condition of the PSA, PSA levels, and age. The data from 36,159 patients were split into two groups, where the first division is known as the training set, and the other division is known as the test set. Data from 25311 patients (70%) was randomly chosen as the training dataset for the ANN, and data from the remaining 10848 patients (30%) was used as the test set. Our network type is a "feed-forward network" and consists of 3 layers: an "input layer", a "hidden layer" and an "output layer" as shown in the following Figure 1.

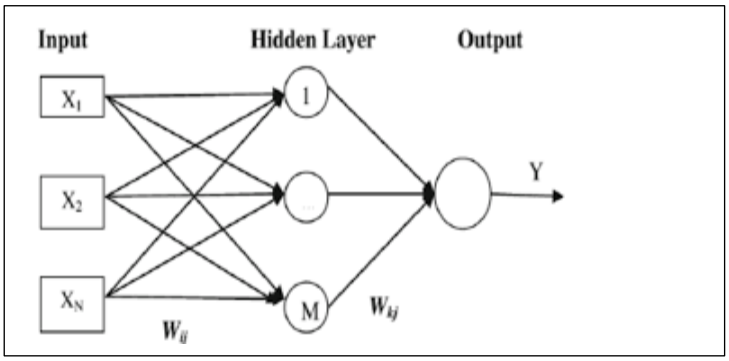

Figure 1: Multi-layer "feed-forward Neural Network" Architecture

The training procedure involved the feed-forward backpropagation algorithm. Various transfer functions are used and tested in the neurons at the unseen and yield layers, including Purlin, Tausig, Logsig, and others. Due to the fact that it gave the most correct results, the Logarithmic-Sigmoid (Logsig) transfer function was chosen as the best option. Based on this, we used TRAINLM and LEARNGDM as training and adaptation learning functions, respectively, to achieve the best result. At the same time, the transfer function is TANSIG. In the first phase, the values of the ANN neuron and bias weights were determined depending on the "training dataset". The repetition of the training at the minimum error point, by reducing the number of neurons to 10 and using 1000000 epochs. The second stage involves the application of the skilled algorithm to the trial dataset, which was processed after completion of the first stage. Part of the training process involved the training of the ANN with more neurons in "hidden layer" (between two and fifty). This was at the initial stage of the training. This was based on the idea of establishing the best method of artificial neural networks that produced the most accurate representation, and it was this method that was selected. Due to this, the ANN scheme, with ten neurons in the "hidden layer", was found to be the best categorisation, which was shown to be effective. The number of epochs that was used was adjusted to test the network structure that was produced in this way. This structure harboured a latent layer which was composed of 10 neurons. The reason has been explained why the most relevant number of the period was used. Brain and performance analysis was developed using a backpropagation algorithm with 10 neurons, which created the best approach containing the fewest number of mistakes. Although the transfer function to be employed in this process is the Tausig transfer function. The calculation of NETj depends on the following equation and other given attributes, such as PSA, PSA level, age, among others.

$NET_j = (W_1)_{i,1}$* PSA + $(W_1)_{i,2}$ * PSA Levels …. + $(W_1)_{i,3}$ *age + ….. +$(W_b)_{i,n}$ *bias

RESULTS AND DISCUSSION

The results of the study prove the possibility of the prediction of prostate cancer by means of the Artificial Neural Network (ANN), which provides an alternative to biopsy for the standard type of cancer. To test the ANN model, some essential metrics were used, and they provide information regarding its evaluation:

The confusion matrix indicated that the ANN model was able to distinguish only 79.8% of the people with prostate cancer from those who are actually not well, and thus can differentiate individuals with the disease. Moreover, the model's accuracy rate in identifying healthy patients was considerable (92.3%), which contributed to an overall success rate of 91.5% in correctly identifying patients. These findings suggest that the ANN model can be effectively used to classify individuals in the accurate diagnosis category.

ROC Curve Analysis: The ROC curve analysis also justifies the efficiency of the ANN model. Values of 0.8490, 0.9126, and 0.9237 for the area AUC were high, indicating that the model performs very well at distinguishing between individuals with and without prostate cancer when applied to training, test, and combined data, respectively. When the AUC is near 1, it indicates that the model can differentiate between the presence of prostate cancer in a person.

Predictive Value: Using predictive values, the analysis indicated that when the ANN model predicted that one would be sick, the prediction and reality aligned with a probability of 73.8% when one was actually ill. On the other hand, there was a 92.0 per cent chance of being healthy when the model indicated being healthy. The values provide information on how well the model performs under various circumstances.

Validation Performance: At an MSE of 0.0644, the best validation performance was realized, which means that the model's

estimates were accurate to the real values provided. Smaller values of MSE usually indicate that the model will perform more accurately.

One of the premises upon which a medical examination is founded is the notion that the outcomes of a patient's tests are a significant factor in determining whether the patient is sick. Through the utilization of the ROC analysis, the actual accuracy of the medical diagnosis outcomes is distinguished. Sensitivity, which refers to the percentage of true positive patients (those who are sick) to all patients who are positive, and specificity, which refers to the percentage of definite negative patients (those who are healthy) to all healthy persons, are the terms that are utilized in this analysis. Both phrases are employed to describe the relationship between the two. We make use of the ROC evaluation, which is a well-known and commonly utilized instrument, to carry out the validation of the identification's sensitivity and specificity. The receiver operating characteristic (ROC) curves will be used to investigate the link between the diagnostic test's sensitivity and specificity to determine the nature of the relationship that would exist between the two features of the test. In accordance with the acronym, the axes of the curves are determined by both true positives (TP) and false positives (FP).

When it comes to positive results, there are two categories: real positives, which are comprised of sick individuals, and false positives, which are comprised of healthy individuals who have been erroneously reported to be sick. In the direction of the boundary between 0 and 1, curves will be inclined to place themselves closest to the boundary. For the purpose of determining whether the test was successful, the curve must be in close contact with both the top border and the coordinate y. On the other hand, the curves give the appearance that the test cannot be successful because the gradient is 45 degrees. This could be since the gradient is steep. It is therefore possible to characterize the performance of the test if the ROC curves are also taken into consideration. In the event that you are conducting a practical test, the AUC ought to be of a size that is enough. Figure 2 gives a depiction of a representation of the receiver operating characteristic (ROC) curve of the “training dataset”. Besides, the curves of the ROC of the test and validation data sets are presented in the figures. The ROC curve of the data set of all patients, as well as that of Figures 3 and 4, has been displayed in Figure 5. The curve of ROC in Figure 4 is also shown in Figure 5.

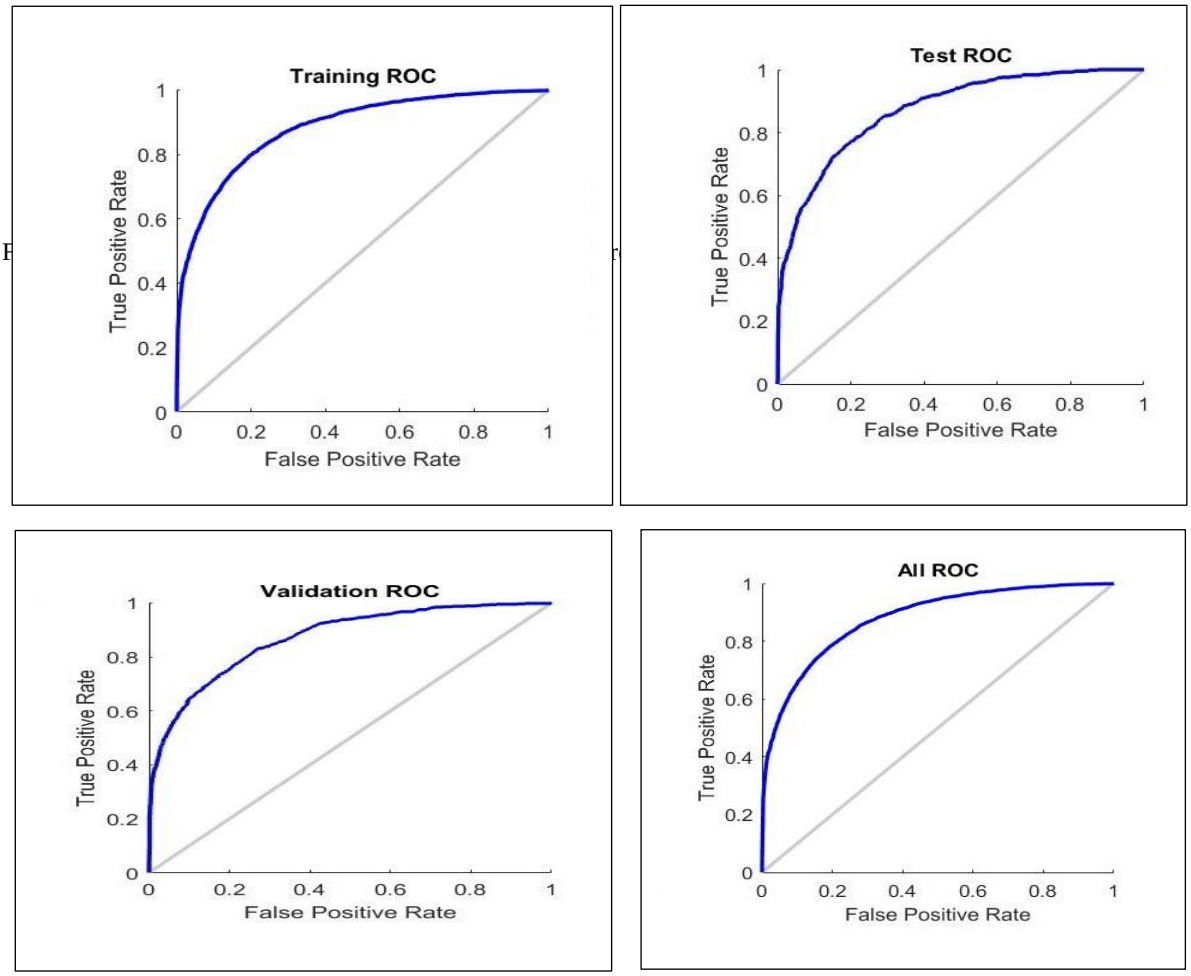

Figure 4: Graph for the ROC analysis of validation data

Figure 5: Graph for the All-ROC analysis of data

According to the analysis of the confusion matrix, it is observed that the distinguishing rate between the sick and the healthy is 79.8%, the rate of recognising the healthy is 92.3%, and the rate of completely correct diagnosis of the ill and the healthy is 91.5%. As such, if the ANN predictive value indicates a sickness judgement, there is a 73.8% possibility of being sick. The ANN predictive value gives a probability that an individual is healthy, which is equivalent to 92.0% in case the ANN predictive value is healthy. The diagram is used to demonstrate this. Based on the interpretation of ROC, 0.8490, in the case of the ANN training

data set (Fig. 2); 0.9126, in the case of the ANN test data set (Fig. 3); 0.9237, in the case of the ANN all data set (Fig. 5), is the area, which remains below the ROC curve. It can be concluded that the research conducted was fruitful in forecasting the potentiality that a diagnosis of prostate cancer will be made because the regions are so large (almost 1).

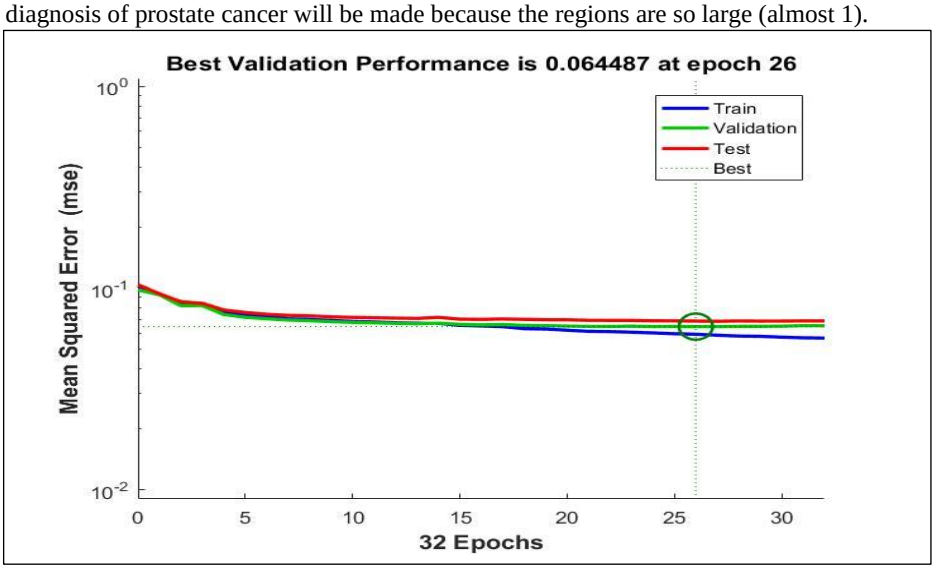

Figure 6: the best validation performance (MSE)

METHODOLOGY:

1. The National Cancer Institute (NCI) data was utilised.
2. ANN model was created with patient age, DRE (Digital Rectal Exam) results, PSA (Prostate-Specific Antigen) results and PSA levels as input variables.
3. The data was separated into training (70) and testing/validation (30) sets.
4. ANN was fed with feed-forward backpropagation algorithm.
5. The model was evaluated by metrics such as confusion matrix, ROC curve analysis, predictive values and mean squared error (MSE).

Key Results & Figures:

Multi-Layer "Feed-forward Neural Network" Architecture is depicted in the Figure 1. The simplest approach to the ANN is a diagram demonstrating the general framework of it: it has an "input layer", a "hidden layer", and an "output layer" in this figure 1. Here X1, X2, and XN are the input data, W are connexions, and Y are the output.

Figures 2, 3, 4, and 5 is representing about the ROC Curve Analysis. These graphs depict the ROC curves of the training and test datasets, the validation and test datasets, and the combined datasets respectively.

The values of the AUC are great:

Training data: 0.8490

Test data: 0.9126

All data: 0.9237

The value of AUC that approximates 1 permits the model to identify the differences between persons with and without prostate cancer. The nearer the curve as shown is to the left of the top corner the model is more predictive.

Figure 6: Optimum validation performance (MSE).

This chart shows the Mean Squared Error (MSE) of the training process.

Optimal validation was attained at epoch 26 with an MSE of 0.0644. The lower MSE demonstrates more accurate model. This graph shows the performance of the model in the prediction of each epoch.

Key Findings are as follows:

The confusion diagram revealed that the model is able to differentiate:

79.8 percent of individuals WITH prostate cancer.

92.3% of healthy individuals.

The total accuracy of diagnosis was 91.5%.

The predictive value implies that when the ANN predicts sickness, the prediction is accurate and the person has a 73.8 percent chance of being sick, and there is a 92.0 percent chance of the prediction being accurate and the person to be healthy when correspondingly the ANN predicts a healthy individual.

The research indicates that ANN models would provide a reasonable accuracy in predicting prostate cancer. The authors however admit that the model is still in need of refinement before it can be used in clinical practise. They suggest that they may add some other variables like lifestyle variables and biopsy variables to enhance the accuracy of the model.

Overall, the research findings demonstrate that prostate cancer can be accurately predicted using an ANN model with a satisfactory level of accuracy. The potential of this model to identify both sick and healthy individuals, as well as its high predictive results and low MSE value, suggest that this model can be adopted as a potential tool for diagnosing prostate cancer. Nonetheless, the authors acknowledge that the model still requires further refinement to be effectively applied in practice, thereby providing high-quality results and reliability.

CONCLUSION

In this research article, the authors investigated the possibility of using an ANN to calculate the occurrence of “prostate cancer” without relying on conventional biopsy procedures. An ANN model was developed, composed of a backpropagation algorithm, and was trained using data comprising levels of PSA, age, and outcomes of digital rectal exams for patients.

The findings demonstrated that the ANN model could be successfully applied in predicting prostate cancer with at least decent precision, as indicated by the confusion matrix and ROC curve analysis. The sensitivity of distinguishing between the truly sick and the sick was 79.8%, whereas the sensitivity of correctly diagnosing the healthy was 92.3%, resulting in a total correct diagnosis rate of 91.5%. There is a high AUC under ROC curve, revealing good predictive ability of the training dataset, test dataset and compound dataset. It produced the best validation performance of 0.0644 at a mean square error (MSE).

Although these are rather optimistic results, the authors acknowledge that the work of the ANN may not be sufficient for direct application in clinical practice. They suggest that additional variables can be included to enhance the model's precision and accuracy, including lifestyle and the results of a biopsy. Prospective studies may be conducted to analyze these factors and compare to confirm whether ANN can be a valuable tool in diagnosing prostate cancer in a non-invasive manner.

REFERENCES

1. Olawuyi, O.; Viriri, S. Deep Learning Techniques for Prostate Cancer Analysis and Detection: Survey of the State of the Art. J. Imaging 2025, 11, 254. https://doi.org/10.3390/jimaging11080254
2. Yuting Cao, Ziyu Sheng, Haibin Zhu, Tingwen Huang, Shiping Wen (2025). Prostate cancer forecasting in small samples based on lightweight neural networks using ensemble learning, Knowledge-Based Systems, Volume 329, Part B,2025,114383, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2025.114383I.
3. Priestley, P., et,al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature, 575(7781), 210–216. https://doi.org/10.1038/s41586-019-1689-y
4. Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. CA a Cancer Journal for Clinicians, 74(1), 12–49. https://doi.org/10.3322/caac.21820
5. Saha A, Bosma JS, et,al.(2024).Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. Lancet Oncol. 2024 Jul;25(7):879-887. doi: 10.1016/S1470-2045(24)00220-1
6. Hendi, A. M., Hossain, M. A., Majrashi, N. A., Limkar, S., Elamin, B. M., & Rahman, M. (2024). Adaptive Method for Exploring Deep Learning Techniques for Subtyping and Prediction of Liver Disease. Applied Sciences, 14(4), 1488. https://doi.org/10.3390/app14041488
7. Czerw, A.; Deptała, A.; Głowacka, M.; Partyka, O.; Pajewska, M.; Czerw, N.; Badowska-Kozakiewicz, A.; Sygit, K.; Kopczyński, Z.; Czarnywojtek, P.; et al. Prostate Cancer—Research Advances in Early Detection. J. Clin. Med. 2025, 14, 3067. https://doi.org/10.3390/jcm14093067
8. Hossain, M. A., Amin, S. T., Eishah, H. A., Hamdan, A. A., Metwally, A. A., & Limkar, S. (2024). Deep Learning for Predictive Modeling of Cardiovascular Disease: Optimizing Neural Networks. In Lecture notes in networks and systems (pp. 371–382). https://doi.org/10.1007/978-981-97-6581-2_30
9. Miao, K.H.; Miao, J.H.; Finkelstein, M.; Chatterjee, A.; Oto, A. Next-Generation Advances in Prostate Cancer Imaging and Artificial Intelligence Applications. J. Imaging 2025, 11, 390. https://doi.org/10.3390/jimaging11110390
10. Web site: https://cdas.cancer.gov/
11. Snow, P. B., Smith, D. S., & Catalona, W. J. (1994). Artificial neural networks in the diagnosis and prognosis of prostate Cancer: a pilot study. The Journal of Urology, 152(5 Part 2), 1923–1926. https://doi.org/10.1016/s0022-5347(17)32416-3.
12. Alhameed, M., Jeribi, F., Elnaim, B. M. E., Hossain, M. A., & Abdelhag, M. E. (2022). Pandemic disease detection through wireless communication using infrared image based on deep learning. Mathematical Biosciences & Engineering, 20(1), 1083–1105. https://doi.org/10.3934/mbe.2023050
13. Islam, R., Imran, A., & Rabbi, M. F. (2024). Prostate Cancer Detection from MRI Using Efficient Feature Extraction with Transfer Learning. Prostate Cancer, 2024, 1–28. https://doi.org/10.1155/2024/1588891
14. R, J., Refaee, et,al (2022). Biomedical image retrieval using adaptive neuro-fuzzy optimized classifier system. Mathematical Biosciences & Engineering, 19(8), 8132–8151. https://doi.org/10.3934/mbe.2022380
15. Mohammed . M. H .T et al., "Enhanced Approach to Predict Early Stage Chronic Kidney Disease," 2024 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS), Manado, Indonesia, 2024, pp. 127-132, doi: 10.1109/AGERS65212.2024.10932874
16. M. M. H. Talukdar et al., "KDAIDL: Kidney Disease Detection Using Artificial Intelligence and Deep Learning," 2024 27th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2024, pp. 3647-3652, doi: 10.1109/ICCIT64611.2024.11022101
17. Takeuchi, T., Hattori-Kato, M., Okuno, Y., Iwai, S., & Mikami, K. (2018). Prediction of prostate cancer by deep learning with multilayer artificial neural network. Canadian Urological Association Journal, 13(5), E145–E150. https://doi.org/10.5489/cuaj.5526
18. Quantrille, T. E., & Liu, Y. (1991). INTRODUCTION TO ARTIFICIAL NEURAL NETWORKS. In Elsevier eBooks (pp. 441–487). https://doi.org/10.1016/b978-0-08-057121-8.50023-3
19. Çınar, M., Engin, M., Engin, E. Z., & Ateşçi, Y. Z. (2008). Early prostate cancer diagnosis by using artificial neural networks and support vector machines. Expert Systems With Applications, 36(3), 6357–6361. https://doi.org/10.1016/j.eswa.2008.08.010

20. Ehsani, Z.; Li, S.; Peng, W. An Intelligent System for the Early Detection of Prostate Cancer. Eng. Proc. 2024, 76, 26. https://doi.org/10.3390/engproc2024076026
21. Sayed-Ahmed, N. M. Z. (2024). Mathematical Modelling and Deep Learning Techniques for Predicting Cardiovascular Disease. Panamerican Mathematical Journal., 34(4), 230–244. https://doi.org/10.52783/pmj.v34.i4.1880
22. Qi, H., Xie, S., Chen, Y., Wang, C., Wang, T., Sun, B., & Sun, M. (2022). Prediction methods of common cancers in China using PCA-ANN and DBN-ELM-BP. IEEE Access, 10, 113397–113409. https://doi.org/10.1109/access.2022.3215706
23. Hossain, M. A., Metwally, A. A., Khan, A. A., Amin, S. T., Hamdan, A. A., & Limkar, S. (2024). Cardiovascular Disease Prediction with Convolutional Neural Networks and Hyperparameter Tuning. In Lecture notes in networks and systems (pp. 357–369). https://doi.org/10.1007/978-981-97-6581-2_29
24. Oh, C.-u.; Kang, H. The Effectiveness and Harms of PSA-Based Prostate Cancer Screening: A Systematic Review. Healthcare 2025, 13, 1381. https://doi.org/10.3390/healthcare13121381
25. M. A. Hossain et al., "IALS: Innovative Approach for Lung Segmentation Applying Artificial Intelligence and Deep Learning," 2025 International Conference on Computer, Electrical & Communication Engineering (ICCECE), Kolkata, India, 2025, pp. 1-8, doi: 10.1109/ICCECE61355.2025.10940152
26. M. Reda. Radwan et. al (2024). Optimizing Electronic Health Records for Enhanced Clinical Decision Support and Improved Patient Care through Advanced Data Integration and Machine Learning Algorithms, Panamerican Mathematical Journal,34(4), pp 282-299, https://doi.org/10.52783/pmj.v34.i4.1884
27. Kania, E.; Janica, M.; Nesterowicz, M.; Modzelewski, W.; Cybulski, M.; Janica, J. Advances and Challenges in Prostate Cancer Diagnosis: A Comprehensive Review. Cancers 2025, 17, 2137. https://doi.org/10.3390/cancers17132137
28. Hendi, A.M., Hossain, M.A., Limkar, S., Majrashi, N.A. (2025). Artificial Intelligence in Precision Oncology, Genomic Medicine, and Disease Monitoring. In: Macha, M.A., Assad, A., Bhat, M.R. (eds) Artificial Intelligence in Human Health and Diseases. Springer, Singapore. https://doi.org/10.1007/978-981-96-8176-1_14
29. Limkar,S; Hossain, M.A;Amin,S.T; Ahmad, Y. (2025). Dynamic Resource Orchestration for Computing, Data, and IoT in Networked Systems: A Data-Centric Approach, Current and Future Cellular Systems: Technologies, Applications, and Challenges,Wiley, 03 January 2025, https://doi.org/10.1002/9781394256075.ch10
30. Reda. M. Radwan, et. al (2024). Developing Secure and Interoperable Health Information Systems Using Blockchain Technology to Enhance Data Privacy, Security, and Accessibility in Healthcare, Panamerican Mathematical Journal, Vol 34, No 4, https://doi.org/10.52783/pmj.v34.i4.1885
31. Ciccone, V.; Garofano, M.; Del Sorbo, R.; Mongelli, G.; Izzo, M.; Negri, F.; Buonocore, R.; Salerno, F.; Gnazzo, R.; Ungaro, G.; et al. Improving Early Prostate Cancer Detection Through Artificial Intelligence: Evidence from a Systematic Review. Cancers 2025, 17, 3503. https://doi.org/10.3390/cancers17213503
32. Uhm, Y.J.; Bang, W.; Chung, J.H.; Oh, C.Y.; Kim, H.; Cho, J.S. Association Between Prostate Cancer Detection Rate and Year of Prostate Biopsy. Life 2025, 15, 260. https://doi.org/10.3390/life15020260