DOI: 10.6977/IJoSI.202406_8(2).0007
K.A. Hernández / J. Systematic Innovation, 8(2), 87-94 (2024)

Prostate cancer prediction using machine learning techniques

Kevin A. Hernández
Research group Automatization and artificial intelligence, Cientek Research Center, Risaralda, Colombia
Corresponding author E-mail: kevin.hernandez.gomez@outlook.com
(Received 10 January 2024; Final version received 27 February 2024; Accepted 13 April 2024)

Abstract

Prostate cancer (PCa) is currently the most frequently diagnosed cancer in men in industrialized nations and ranks as the second leading cause of male cancer-related deaths globally, early detection is crucial. Originating in the walnut-shaped gland beneath the bladder, PCa poses a significant risk when not identified in its early stages. The diagnostic process, requiring expertise from radiologists, pathologists, and physicians, is time-consuming and introduces variability, potentially leading to delayed or incorrect diagnoses. This underscores the need for efficient and reliable diagnostic tools in addressing the escalating challenge of PCa diagnosis. This study addresses the critical challenge of PCa diagnosis by employing a comprehensive approach involving feature selection methods and model performance evaluation. Utilizing a PCa dataset from Kaggle, consisting of 100 patient observations with eight independent features and a binary diagnosis result, the study explores the nuanced nature of feature relevance in PCa classification. Comparative analyses of Principal Component Analysis (PCA) and ReliefF feature selection methods reveal the limitations of PCA's emphasis on a dominant feature, while ReliefF, incorporating a distributed set of features, demonstrates improved model accuracy and stability. The Random Forest (RF) model, selected through meticulous parameter tuning, achieves an impressive 95% accuracy by leveraging a substantial number of estimators, limited tree depth, and balanced sample splitting. The findings underscore the crucial interplay between feature selection methods and model parameters in optimizing the accuracy and reliability of PCa classification models. Given the anticipated rise in PCa incidence, this research contributes valuable insights for enhancing diagnostic efficiency and addressing the challenges posed by traditional diagnostic procedures.
Keywords: Prostate Cancer, Machine Learning, Feature selection.

1. Introduction

At present, prostate cancer (PCa) stands as the most commonly diagnosed malignancy among men in highly industrialized nations and ranks as the second primary cause of male cancer-related fatalities globally. As the population continues to expand and age, it is anticipated that the global incidence of PCa will rise, reaching nearly 2.4 million new cases annually by the year 2040 (De Vos et al., 2023). PCa originates in the compact, walnut-shaped gland located beneath the bladder and in front of the rectum. When not identified in its early stages, PCa can pose a considerable risk, leading to a notable fatality rate. According to a 20-year actuarial cumulative estimate, the likelihood of death from prostate cancer is significant (ACS, 2023). Furthermore, the clinical procedures for diagnosing prostate cancer (PCa) necessitate considerable time and expertise from radiologists, pathologists, and physicians. They meticulously assess and assign a grade or stage before considering treatment options based on factors such as the cancer stage, severity, and other relevant considerations. Unfortunately, the routine diagnostic process relies on human intervention, introducing variability in outcomes that may result in delayed or incorrect diagnoses (Gravade et al., 2023).

Additionally, the diagnosis of PCa continues to be challenging because each cascade element is not fully replicated in the metastasis of prostate cancer. Traditional methods such as the digital rectal test (DRE), prostate-specific antigen (PSA) blood test, and ultrasonography are employed for PCa detection. However, these methods exhibit low sensitivity and specificity, falling short of meeting medical standards (Naeem et al., 2023).

On the other hand, machine learning (ML) algorithms have proven effective in identifying gene biomarkers associated with PCa. Researchers are drawn to this technology because of its ability to uncover hidden patterns in the data and extract relationships between features using a set of mathematical rules and statistical assumptions (Chen et al., 2022).

In this paper we present a methodology for the PCa diagnosis system based on ML classifier, the aim of this study is to use and compare various supervised machine learning algorithms like Multilayer Perceptron (MLP), Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), Naïve Bayes (NB) and Random Forest (RF). The remaining sections of the paper are organized as follows: Section 2 introduces the related works in literature. In Section 3, the materials and methods are presented. Section 4 discusses the experimental results and findings. Finally, Section 4 provides the conclusion for the paper.

2. Literature review

The examination of ML algorithms in the context of predicting PCa represents a central theme in current medical research. With a primary objective of improving the survival rates of individuals diagnosed with PCa, the development of robust prediction models holds paramount importance, for example in (Molla et al., 2023) the exploration is undertaken by utilizing a variety of ML techniques, namely SVM, KNN, NB, RF, and Logistic Regression (LR) algorithms. The objective is to predict PCa outcomes with greater precision. Notably, among the diverse ML techniques investigated, LR emerges as particularly promising, showcasing a noteworthy 86.21% accuracy in prediction results. These findings underscore the potential applicability of LR as a reliable tool for PCa prediction.

On the other hand, in (Laabidi, & Aissaoui, 2020) they specific focus on the study involves predicting diabetes and PCa, utilizing eight distinct machine learning architectures. The experiments conducted reveal promising results, with an overall accuracy of 81.3% for PCa diagnosis. Notably, the Recurrent Neural Network (RNN) emerged as the top-performing model, showcasing superior accuracy compared to other architectures. However, LR demonstrated noteworthy results, particularly when applied to scaled features.

Moreover, the methodology built in (Araujo et al., 2023) upon a comprehensive analysis of various clinical variables extracted from patients' medical records, including age, race, diabetes mellitus, alcoholism, smoking, systemic arterial hypertension, digital rectal examination, and total prostate-specific antigen levels. To validate the efficacy of the method, machine learning algorithms such as SVM, NB, KNN, DT, and MLP were employed. These algorithms were utilized to predict the likelihood of PCa presence or absence based on the gathered clinical data. The evaluation of the method's performance employed an accuracy metric, with the Linear SVM model exhibiting the highest accuracy at 86.8%.

3. Materials and methods

In this section, we present a thorough overview of the methodology utilized, feature selection techniques, covering machine learning algorithms, and a detailed description of the dataset.

3.1. Feature selection methods

Principal Component Analysis (PCA): It is a procedure leveraging statistical methods to derive features from a dataset. This involves determining the eigenvalues of various features within the dataset and projecting them into a lower-dimensional space. The derived features are commonly known as principal components. Despite being sensitive to missing or outlier values, PCA aims to preserve minimal dimensionality while retaining valuable or essential information (Alhanaya et al., 2023).

ReliefF: The fundamental concept behind the Relief algorithm is to assess features by their effectiveness in distinguishing instances that are in close proximity to each other. For every selected instance, the algorithm identifies its two nearest neighbors: one belonging to the same class, termed the nearest hit, and the other from a different class, referred to as the nearest miss. This algorithm assigns higher weights to features that effectively differentiate instances from diverse classes. Similarly, the ReliefF algorithm is developed on the same underlying rationale (Yong & Gao, 2023).

3.2. Machine learning algorithms

Multilayer Perceptron (MLP): It is a soft computing tools for constructing reliable models to address diverse and intricate engineering problems, mimicking the structure of biological neural networks. The architecture mainly consists of three components: an input layer containing features, hidden layers with synapses, a summing point, and an activation function, and an output layer displaying results. This network configuration is

commonly referred to as MLP, employing multiple perceptron's or neural network units to compute specific input data. Each layer, depending on the input elements from preceding layers, features a definite number of nodes or neurons interconnected by synapses or weights converging at the summation point, resulting in a modified signal post-multiplication by varying weights. The summation point combines input signals linearly, potentially yielding a substantial output amplitude. To constrain the signal amplitude from the summation point, an activation function is employed (Deka et al., 2023).

Support Vector Machines (SVM): Is a supervised learning algorithm suitable for classification and regression. In a classification scenario, it separates labeled training data into positive and negative classes within an n-dimensional space. The SVM's objective is to identify a hyperplane that maximizes the distance between the plane and the nearest data points, known as the maximal margin hyperplane. The hyperplane's parameters, such as the n-dimensional weight vector and bias value, are determined during the learning phase. If the data are not linearly separable, the algorithm allows for some misclassification using slack variables and an error penalty parameter. The optimal hyperplane is defined by solving a convex quadratic optimization problem. In a visual representation, support vectors represent the nearest data points to the hyperplane, and the distance between them constitutes the margin. This approach is effective in addressing diverse classification challenges in engineering problems (Araste et al., 2023).

K-Nearest Neighbor (KNN): a supervised learning method addressing grouping problems, stands out as one of the most commonly employed classification algorithms in literature. Its classification process relies on known class data, making it an example-based algorithm learning from the training set. Despite its simplicity, KNN consistently delivers competitive results and can even outperform more complex learning algorithms in certain cases, especially when dealing with a smaller number of classes. This method proves to be a simpler yet effective machine learning approach, particularly in situations with multiple categorized data points. KNN finds applicability not only in classification but also in solving regression problems, especially when independent variables are quantitative, and the classification process depends on the distances between observations. Although possessing a straightforward structure, KNN does entail a high computational cost (Erdem & Bozkurt., 2021).

Decision Tree (DT): Is a formalism used to express mappings, comprising tests or attribute nodes connected to two or more subtrees and leaf nodes. Decision nodes or leaf nodes are labeled with a class, representing the decision. A test node calculates an outcome based on the attribute values of an instance, with each possible outcome associated with one of the subtrees. Classification of an instance involves starting at the root node of the tree. If this node is a test, the outcome for the instance is determined, and the process continues using the appropriate subtree. When a leaf is encountered, its label provides the predicted class for the instance (Podgorelec, 2002).

Naïve Bayes (NB): Serves as a straightforward probability classifier, determining probabilities by tallying the frequency and combinations of values within a given dataset. Employing Bayes's theorem, the algorithm operates under the assumption of independence among all variables, given the class variable. While this conditional independence assumption may be deemed "naive" and is rarely valid in real-world applications, the algorithm demonstrates a rapid learning capability across diverse controlled classification problems. Bayes's theorem, a mathematical formula named after the 18th-century British mathematician Thomas Bayes, is utilized to calculate conditional probability (Saritas & Yasar, 2019).

Radom Forests (RF): Is a combination of classifiers where each classifier contributes a single vote for assigning the most frequent class to the input vector. The majority vote in RF representing the class prediction of the random forest tree. The amalgamation of many classifiers gives RF distinct characteristics, setting it apart from traditional DT. Therefore, RF should be perceived as a novel concept in classifiers. RF enhances tree diversity by growing them from different training data subsets, which are created through bagging or bootstrap aggregating. Bootstrap aggregating involves randomly resampling the original dataset with replacement, thus generating subsets with varied data. RF serves as an ensemble classification algorithm that utilizes trees as base classifiers (Rodriguez-Galiano., et al 2012).

3.3. Dataset description

To assess the performance of the methods evaluated in this study, a prostate cancer dataset was employed in the initial phase of the design flow. The dataset, accessible through the Kaggle platform (Sajid, 2018),

DOI: 10.6977/IJoSI.202406_8(2).0007
K.A. Hernández / J. Systematic Innovation, 8(2), 87-94 (2024)

comprises observations from 100 patients (62 records for PCa patients and 38 records for non-PCa patients). It includes eight independent features (radius, texture, area, perimeter, compactness, smoothness, fractal dimension, symmetry) and one dependent variable (diagnosis result). The label is represented by the diagnosis results, categorized by two values "B" for benign tumors and "M" for malignant tumors. The detailed information about the data is presented in Table 1

Table 1. Prostate cancer dataset feature descriptions.

Feature	Description
Radius	Refers to the average distance from the center to the perimeter of the cancer cell. This feature is often used to characterize the size of the cell.
Texture	Describes the variation in gray-scale intensity of the cancer cell, providing information about the homogeneity of the cell's internal structure.
Perimeter	Represents the total length of the boundary of the cancer cell, offering insights into the shape and contour.
Area	Denotes the total area covered by the cancer cell, contributing to the overall size assessment.
Smoothness	Describes the local variation in radius lengths, giving an indication of how smooth or irregular the cancer cell surface is.
Compactness	Reflects the compactness of the cancer cell shape, derived from the ratio of perimeter^2 to area.
Symmetry	Represents the symmetry of the cancer cell shape, providing information about its regularity.
Fractal dimension	Describes the complexity of the cancer cell shape at different scales, offering insights into the irregularity and intricacy of the cell structure.

4. Results and discussion

In this section, we unveil the outcomes derived from the applied methodology. We start with the results of the feature and model selection stage; wherein numerous experiments were conducted by training each model under different feature selection configurations (None, PCA and ReliefF). The optimal model from this stage was then chosen for subsequent parameter tuning.

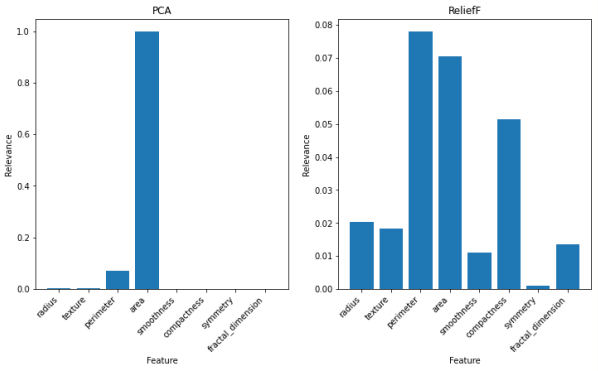


Figure 1. Feature relevance, PCA method (left) and ReliefF (right).

For the PCA method, the feature relevance score for "area" is the highest, close to 1.0, while the other features have near to zero scores. This suggests that "area" is the dominant feature that explains most of the variance in the data, and the other features are redundant or irrelevant. However, this does not mean that "area" is the best feature for classification or regression, as it may not capture the differences between the classes.

In the other way, the ReliefF method, the feature relevance scores are more distributed among the features. The feature "perimeter" has the highest score, but it is much lower than in the PCA method. This indicates that "perimeter" is still a relevant feature for classification, but it is not the only one. Other features, such as "area" and "compactness", also have notable scores, which means that they are also useful for distinguishing between the classes. The remaining features have relatively low relevance scores, which means that they are less important or redundant.

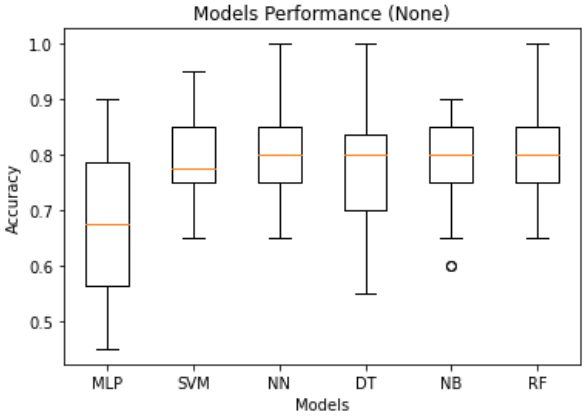


Figure 2. Models' performance for None feature selection.

The figure reveals that among the models, MLP has the widest range of accuracy (0.68±0.11), which suggests that it is unstable and sensitive to the data. SVM, DT and NB have more consistent accuracy, but they are still below 0.8, 0.79±0.08, 0.76±0.1 and 0.78±0.08 respectively. NN, and RF have similar median accuracy and variability, 0.81±0.08 and 0.82±0.08 respectively. These results imply that as some features may be irrelevant or redundant for the classification task.

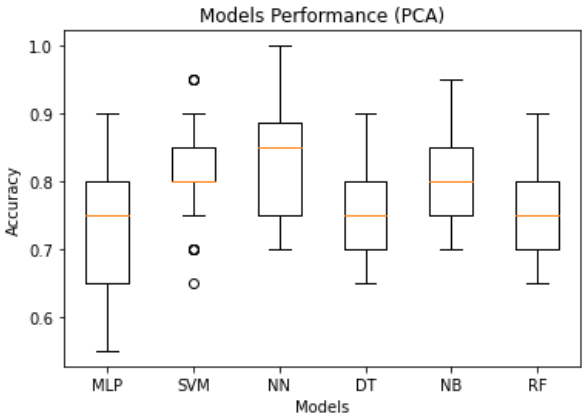


Figure 3. Models' performance for PCA feature selection.

The figure shows the PCA feature selection based on the "area" as unique feature. Equal to none feature selection, MLP has the widest range of accuracy (0.74±0.09), similar to DT and RF with 0.76±0.07 each. NN has the highest median accuracy 0.83±0.07 near are SVM and NB with 0.82±0.07 each. This implies that PCA feature selection may not be the best choice for this classification task, as it only considers the "area" feature and ignores the other features that may be relevant or informative.

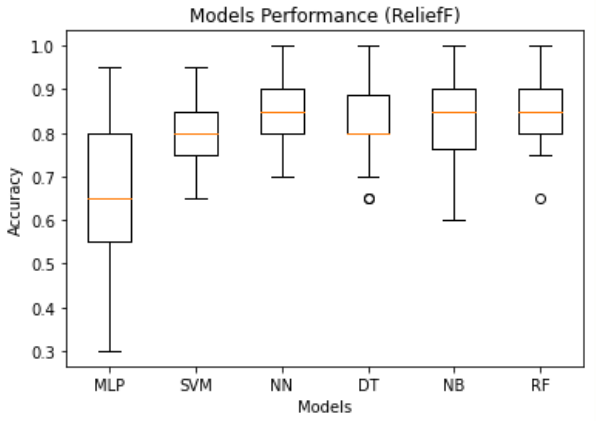


Figure 4. Models' performance for ReliefF feature selection.

DOI: 10.6977/IJoSI.202406_8(2).0007
K.A. Hernández / J. Systematic Innovation, 8(2), 87-94 (2024)

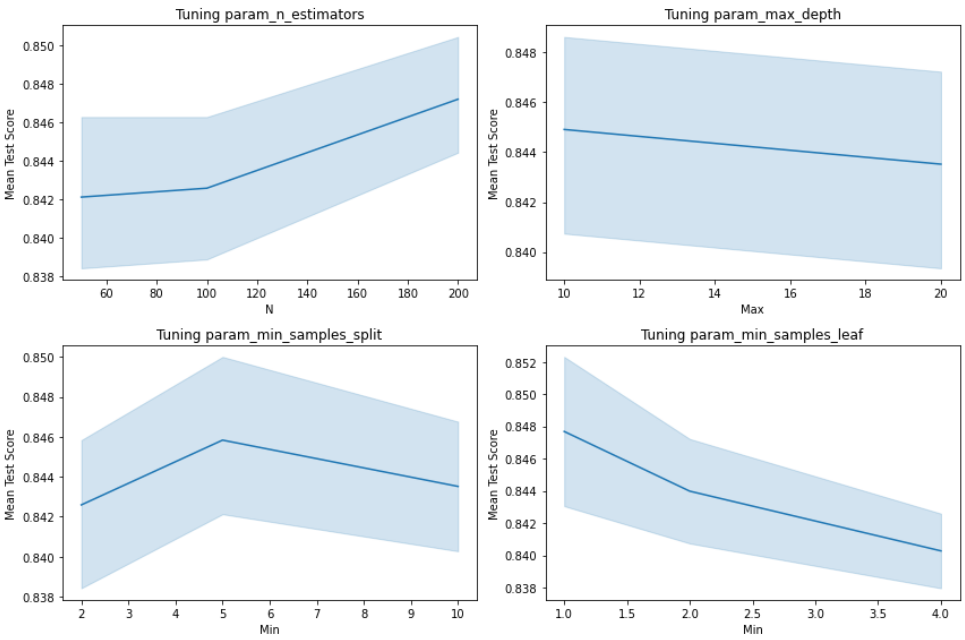


Figure 5. RF parameter tunning values.

ReliefF feature selection is based on three features: “perimeter,” “area,” and “compactness”. Here the results improve for four of the six models (NN, DT, NB, and RF) with 0.84±0.07, 0.82±0.08, 0.83±0.08 and **0.85±0.08** respectively, being RF the best model in all experiments, while MLP and SVM give lower scores 0.65±0.15 and 0.81±0.08. This implies that ReliefF feature selection may be beneficial for improving the accuracy and stability of the models, as it considers the “perimeter”, “area” and “compactness” that are relevant and good predictors for prostate cancer classification tasks.

Finally, we opted for the RF model during the parameter tuning phase. For this stage, we employed a parameter grid encompassing "number of estimators" (50, 100, 200), "max depth" (None, 10, 20), "minimum samples split" (2, 5, 10), and "minimum samples leaf" (1, 2, 4). The optimal model was determined to have the following parameter values: 200 for the number of estimators, 10 for max depth, 5 for minimum samples split, and 1 for minimum samples leaf (**Figure 5**). This refined model achieved an impressive accuracy score of 95% for diagnosing PCa.

The noteworthy accuracy of this model can be attributed to the combination of a substantial number of estimators (200), an appropriately limited tree depth (10), and a balanced approach to splitting samples. The minimal leaf samples (1) further contribute to the model's precision, ensuring that each leaf node captures a sufficient amount of information without overfitting the data. This comprehensive parameter selection enables the Random Forest model to robustly discern patterns in the dataset, leading to its high accuracy in PCa diagnosis.

5. Conclusion

In conclusion, the comparative analysis of feature selection methods and subsequent model performance evaluation highlights the nuanced nature of feature relevance in PCa classification. PCA, with its emphasis on the dominant "area" feature, may oversimplify the task by neglecting other informative features, potentially compromising classification accuracy. In contrast, ReliefF, incorporating "perimeter," "area," and "compactness," demonstrates improved model accuracy and stability, emphasizing the significance of a more distributed feature selection approach. The RF model, chosen through parameter tuning, achieves an impressive 95% accuracy by effectively leveraging a substantial number of estimators, limited tree depth, and balanced sample splitting. This underscores the importance of a meticulous parameter selection process, contributing to the model's robust ability to discern meaningful patterns in the PCa dataset. Overall, the study underscores the critical interplay between feature selection methods and model parameters in

optimizing the accuracy and reliability of PCa classification models.

References

ACS (American Cancer Society). (2023). Survival Rates for Prostate Cancer. Atlanta, GA, USA.

Alhanaya, M., & Ateyeh Al-Shqeerat, K. H. (2023). Performance Analysis of Intrusion Detection System in the IoT Environment Using Feature Selection Technique. Intelligent Automation & Soft Computing, 36(3).

Araste, Z., Sadighi, A., & Jamimoghaddam, M. (2023). Fault diagnosis of a centrifugal pump using electrical signature analysis and support vector machine. Journal of Vibration Engineering & Technologies, 11(5), 2057-2067.

Araujo, W. B., Santana, E. E., Sousa, N. P., Junior, C. M., Allan Filho, K. D. B., Moura, G. L., ... & Silva, F. C. (2023). Method to aid the diagnosis of prostate cancer using machine learning and clinical data.

Chen, S., Jian, T., Chi, C., Liang, Y., Liang, X., Yu, Y., ... & Lu, J. (2022). Machine learning-based models enhance the prediction of prostate cancer. Frontiers in Oncology, 12, 941349.

De Vos, I. I., Luiting, H. B., & Roobol, M. J. (2023). Active Surveillance for Prostate Cancer: Past, Current, and Future Trends. Journal of Personalized Medicine, 13(4), 629.

Deka, M. J., Kalita, P., Das, D., Kamble, A. D., Bora, B. J., Sharma, P., & Medhi, B. J. (2023). An approach towards building robust neural networks models using multilayer perceptron through experimentation on different photovoltaic thermal systems. Energy Conversion and Management, 292, 117395.

Erdem, E., & Bozkurt, F. (2021). A comparison of various supervised machine learning techniques for prostate cancer prediction. Avrupa Bilim ve Teknoloji Dergisi, (21), 610-620.

Gavade, A. B., Nerli, R., Kanwal, N., Gavade, P. A., Pol, S. S., & Rizvi, S. T. H. (2023). Automated diagnosis of prostate cancer using mpmri images: A deep learning approach for clinical decision support. Computers, 12(8), 152.

Laabidi, A., & Aissaoui, M. (2020, April). Performance analysis of Machine learning classifiers for predicting diabetes and prostate cancer. In 2020 1st international conference on innovative research in applied science, engineering and technology (IRASET) (pp. 1-6). IEEE.

Molla, M. I., Jui, J. J., Rana, H. K., & Podder, N. K. (2023, January). Machine Learning Algorithms for the Prediction of Prostate Cancer. In Proceedings of International Conference on Information and Communication Technology for Development: ICICTD 2022 (pp. 471-482). Singapore: Springer Nature Singapore.

Naeem, A., Khan, A. H., u din Ayubi, S., & Malik, H. (2023). Predicting the Metastasis Ability of Prostate Cancer using Machine Learning Classifiers. Journal of Computing & Biomedical Informatics, 4(02), 1-7.

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision trees: an overview and their use in medicine. Journal of medical systems, 26, 445-463.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS journal of photogrammetry and remote sensing, 67, 93-104.

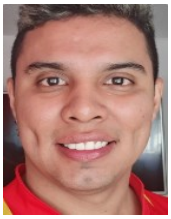
Sajid S. (2018). Prostate cancer dataset, [Online]. Available: https://www.kaggle .com/sajidsaifi/prostate cancer

Saritas, M. M., & Yasar, A. (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. International journal of intelligent systems and applications in engineering, 7(2), 88-91.

Yong, X., & Gao, Y. L. (2023). Improved firefly algorithm for feature selection with the ReliefF-based initialization and the weighted voting mechanism. Neural Computing and Applications, 35(1), 275-301

DOI: 10.6977/IJoSI.202406_8(2).0007
K.A. Hernández / J. Systematic Innovation, 8(2), 87-94 (2024)

AUTHOR BIOGRAPHY



Kevin A. Hernández has been a Researcher at Cientek Research Center in Colombia since 2023. Kevin graduated from physics engineering and his master's degree in electrical engineering at Universidad Tecnologica de Pereira. He is currently aiming for his Ph.D. His areas of interests include Machine Learning and Deep Learning.